Noname manuscript No.
(will be inserted by the editor)

Robust Statistical Methods for Empirical Software Engineering

Barbara Kitchenham · Lech Madeyski (✉) · David Budgen · Jacky Keung · Pearl Brereton · Stuart Charters · Shirley Gibbs · Amnart Pohthong

Received: date / Accepted: date

Abstract Context: There have been many changes in statistical theory in the past 30 years, including increased evidence that non-robust methods may fail to detect important results. The statistical advice available to software engineering researchers needs to be updated to address these issues.
Objective: This paper aims both to explain the new results in the area of robust analysis methods and to provide a large-scale worked example of the new methods.
Method: We summarise the results of analyses of the Type 1 error efficiency and power of standard parametric and non-parametric statistical tests when applied to non-normal data sets. We identify parametric and non-parametric methods that are robust to non-normality. We present an analysis of a large-scale software engineering experiment to illustrate their use.

Barbara Kitchenham
School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK. E-mail: b.a.kitchenham@keele.ac.uk

Lech Madeyski (Corresponding Author)
Faculty of Computer Science and Management, Wroclaw University of Science and Technology, Poland. E-mail: Lech.Madeyski@pwr.edu.pl `http://madeyski.e-informatyka.pl/`

David Budgen
School of Engineering & Computing Sciences, Durham University, Durham, UK. E-mail: david.budgen@durham.ac.uk

Jacky Keung
Hong Kong City University, HK, E-mail: jacky.keung@cityu.edu.hk

Pearl Brereton
School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK. E-mail: o.p.brereton@keele.ac.uk

Stuart Charters and Shirley Gibbs
Department of Applied Computing, Lincoln University, NZ. E-mail: {stuart.charters;shirley.gibbs}@lincoln.ac.nz

Amnart Pohthong
Prince of Songkla University, Thailand. E-mail: amnart.p@psu.ac.th

Results: We illustrate the use of kernel density plots, and parametric and non-parametric methods using four different software engineering data sets. We explain why the methods are necessary and the rationale for selecting a specific analysis.
Conclusion: We suggest using kernel density plots rather than box plots to visualise data distributions. For parametric analysis, we recommend trimmed means, which can support reliable tests of the differences between the central location of two or more samples. When the distribution of the data differs among groups, or we have ordinal scale data, we recommend non-parametric methods such as Cliff's δ or a robust rank-based ANOVA-like method.

Keywords empirical software engineering · statistical methods · robust methods · robust statistical methods

1 Introduction

In 1996, the first author of this paper wrote a book on software metrics (Kitchenham, 1996). In the book chapter addressing statistical methods, her advice was to use box plots to visualize data. Box plots are based on the median and fourth statistics (which are similar to quartiles), so are more robust than any graphics based on means. If data were non-normal, she advised the use of non-parametric methods such as Kruskal-Wallis rank tests to compare multiple samples. With more complicated designs she advised using analysis of variance methods (ANOVA) with transformations if necessary.

Other software engineering researchers preferred to avoid the non-parametric tests relying on the Central Limit Theorem, which proves that for any set of N identically distributed variables, the mean of the variable values will be approximately normal, with mean, μ, and variance, σ^2/N. The Central Limit Theorem provides the justification for use of methods based on the normal distribution to handle small samples, such as t-tests. Their choice was justified by the observation that simulation studies had suggested the t-test and ANOVA were quite robust even if some of the variances within groups differed (Box, 1954).

In this paper, we discuss more recent studies of the t and F tests that show that if data sets are not normal (that is the data sets do not originate from a Gaussian distribution), the statistical tests may not be trustworthy. Statistical hypothesis testing can make two kinds of error. Type I errors occur when we reject the null hypothesis when it is in fact true, which is also called a false positive. Conventionally statisticians choose a probability level they believe is acceptable for a Type I error, which is referred to as the α-level. It is usually set to values of 0.05 or 0.01. Type II errors occur when we fail to reject the null hypothesis when it is in fact false, which is also called a false negative. Statisticians usually prefer the probability of a Type II, which is referred to as the β-level to be 0.2 or less. A related concept is statistical power which is the probability of correctly rejecting the null hypothesis, so that $power = 1-\beta$. Although the probability of either type of error is decreased by

using larger sample sizes, aiming for a very low α-level given a predetermined sample size will increase the achieved β-level and reduce power. Studies of classical statistical tests under conditions of non-normality have shown that the assumed α levels of tests are likely to be incorrect, and the power of various tests may be unacceptably low.

In a study of 440 large-sample achievement and psychometric measures data sets, Micceri (1989) found all to be significantly non-normal. He noted that data values were often discrete, while distributions exhibited skewness, multiple modes, long tails, large outlying values and contamination. In our experience, similar issues affect software engineering data sets[1]. The prevalence of non-normal data sets and recent studies showing poor performance of classical statistical tests on such data sets, suggest that empirical software engineers need a major re-think of the techniques used for statistical analysis. Recent statistical studies have not only identified analysis problems, they have also introduced methods of addressing these problems. In this paper we identify a number of robust methods that address the problems associated with non-normal data.[2]

Interest in robust methods dates back to the early 1960's, when Tukey and his colleagues introduced the concept of Exploratory Data Analysis (EDA), see for example Mosteller and Tukey (1977) or Hoaglin et al (1983). Tukey and colleagues pointed out that although classical statistical techniques are designed to be the best possible analysis methods when specific assumptions apply:

> "... experience and further research have forced us to recognize that classical techniques can behave badly when the practical situation departs from the ideal described by such assumptions."

Behrens (1997) summarises EDA as involving:

- An emphasis on understanding the data using graphic representations of the data.
- A focus on tentative model building and hypothesis generation as opposed to confirmatory analysis.
- Use of robust measures.
- Positions of skepticism and flexibility regarding which techniques to apply.

Tukey and his colleagues introduced graphical techniques such as box plots and stem-and-leaf displays and emphasized the importance of residual analysis. All these methods are well known to empirical software engineering researchers. However, they were also concerned with the construction of new types of measures (see Section 3.2 and Appendix A), which have not been taken up by software engineering researchers.

[1] There has not been a systematic review of all publicly available software engineering data sets. However, Whigham et al (2015) propose the use of the logarithmic transformation for their proposed cost estimation baseline, and suggest that non-Normality is the norm for cost estimation data sets.

[2] This part of the paper is based on a keynote paper given at the EASE-2015 conference (Kitchenham, 2015).

They emphasized using robust and resistant methods that can be regarded as optimal for a broad range of situations. They proposed the following definitions:

- *Resistant measures and methods* are those that provide insensitivity to localized misbehavior in data. Resistant methods pay attention to the main body of the data and little to outliers.
- *Robust methods* are those that are insensitive to departures from assumptions related to a specific underlying model.

In this paper we focus on robust measures and robust methods and regard resistance as being a property of such measures and metrics.

Tukey and his colleagues preferred robust and resistant methods to non-parametric methods. They point out that distribution-free methods treat all distributions equally, but robust and resistant methods discriminate between those that are more plausible and those that are less plausible. To distinguish their approaches from classical methods, they introduced new terms such as *batch* as an alternative to *sample* and *fourths* as opposed to *quartiles*. Currently few of these terms are still in use with the exception of fourths, which are used in the context of box plots. In this paper we will introduce methods that arose from EDA concepts (specifically central location measures related to the median and trimmed means) but will also emphasize the use of *robust* non-parametric methods as viable alternatives to parametric analysis. An important issue raised in this paper is that under certain conditions non-parametric rank-based tests can themselves lack robustness.

We illustrate the new methods using software engineering data and analyse the results of a large scale experiment as an example of the use of these techniques. However, before considering the robust analysis methods, we introduce the use of kernel density plots as a means of visualising data. These can provide more information about the distribution of a data set than can be obtained from box plots alone.

Other researchers have started to adopt the robust statistical methods discussed in this paper, e.g., Arcuri and Briand (2011), El-Attar (2014), Madeyski et al (2014, 2012). In particular, Arcuri and Briand (2014) have undertaken an important survey of statistical tests for use in assessing randomized algorithms in software engineering. We agree with many of their recommendations (particularly their preference for non-parametric methods), but, in this paper, we focus on approaches suitable for relatively small samples such as those obtained from human-based experiments, or algorithms that give rise to relative small data sets (such as project cost estimation models), rather than the large data sets they discuss. The main contribution of this paper is to provide an overview of the techniques with extended examples of their use and an introduction to the underlying theory. In addition, based upon using the open source **R** statistical programming language (R Core Team, 2015), the **reproducer R** package by Madeyski (2015) complements this paper, as well as the paper by Jureczko and Madeyski (2015), with the aim of making our work *reproducible* by others (Gandrud, 2015). All of our data sets are encapsulated

in the **reproducer R** package we have created and made available from CRAN – the official repository of **R** packages. All of the figures in the paper (except the figures in Appendices A and B, which do not depend on data sets collected by us) are built on the fly from data sets stored in the **reproducer** package.

2 Problems with conventional statistical tests

In this section we summarise the results of studies that have investigated the performance of parametric and non-parametric statistical tests under conditions of non-normality. These studies identify some of the problems that can occur when using conventional statistical tests on data exhibiting characteristics found in real data sets.

2.1 Parametric tests

Student's t distribution was intended as a small sample correction for normal data, so it is necessary to consider what happens if the population is not normal. We consider first the one-sample case where we want to put confidence limits on the sample mean. With a lognormal distribution (i.e., a skewed distribution with a long tail but relatively few outliers), Wilcox and Keselman (2003) report that with sample size $n = 20$, the actual distribution varies considerably from the t-distribution. Furthermore, the problem persists even when $n = 160$. In this case, using an alpha value equal to 0.1:

- The lower tail probability of a Type I error is 0.11 rather than 0.05.
- The upper tail probability of a Type I error is 0.02 rather than 0.05.

Hence, in this case, the actual probability of a Type I error is 0.13 instead of 0.1. With $n = 200$, the lower tail Type I error is 0.07 instead of 0.05.

Wilcox and Kesleman also investigated what would happen if the distribution was skewed *and* had *heavy tails* (i.e., a relatively large number of outliers). In this case, with $n = 20$ and a normal distribution, there is a .95 probability that t will be between -2.09 and 2.09 but the actual distribution based on 5000 samples, had 0.025 and 0.975 quantiles of -8.5 and 1.29 respectively. With $n = 300$, the quantiles were -2.50 and 1.70 compared with theoretical values (under normality) of -1.96 and 1.96 respectively.

There are also problems with "contaminated" normal distributions where the majority of the data comes from one distribution and a small percentage of the data comes from a distribution with a much larger variance. In this case, the variance is larger than the uncontaminated distribution, which means that the standard deviation is relatively large and the presence of the outliers that cause the variance inflation may be masked. Variance inflation will also increase the likelihood of Type II errors.

In the two-sample case, if, the two groups exhibit the same amount of skewness and sample sizes are equal, the t test should perform correctly

because the difference between the mean values should be distributed symmetrically. However, empirical studies summarised by Wilcox (2012) confirm that if distributions vary in shape, Type I errors may be incorrect.

If the variance is different in each group (i.e., the data exhibit heteroscedasticity), Ramsey (1980) showed that the t test is robust if:

- Group sizes are equal.
- Data in each group are normally distributed.
- Sample sizes are not small, where small was defined as a sample size of $n < 15$ in each group.

Box (1954) reported acceptable behaviour of the t test under heteroscedasticity (unequal variances), but his study restricted the extent of the difference between the variances. The maximum heteroscedasticity he studied was one variance being three times larger than the other.

In contrast, Wilcox (1998) found problems with Type I and Type II errors with heteroscedastic data if:

- Data were normal and sample sizes were unequal for two or more groups.
- Data were normal, sample sizes were the same and there were four or more groups.
- Data were non-normal when comparing two or more groups even if sample sizes were equal.

Thus, recent studies imply that:

- We need large sample sizes to avoid problems with non-normal data.
- With small samples and non-normal data, t tests might be very problematic.
- Data distributions exhibiting combinations of non-normal properties usually have more severe problems than distributions with only one non-normal property.
- Except under specific conditions, the classical parametric t and F tests are vulnerable to non-normality and heteroscedasticity.

Overall the problem is that, although the Central Limit theory confirms that (under most practical situations) the mean of a sample is distributed normally, there are no such guarantees about the variance of a sample. With messy data sets, estimates of the variance may be far from reliable, rendering unreliable any statistical tests, such as the t test, that rely upon knowing the variance of a mean value.

2.2 Non-parametric tests

Given that there might be problems with parametric tests, what about the non-parametric methods? Unfortunately, simulation studies have shown that the large sample approximation for the Mann-Whitney-Wilcoxon (MWW) tests and Kruskal-Wallis test are strongly affected by unequal variances, even if

sample sizes are equal. In fact they can be less robust than the standard t test, see (Zimmerman and Zumbo, 1993; Zimmerman, 2000).

Furthermore, problems with the rank-methods can affect the results of statistical packages and can make the difference between finding a significant result and finding a non-significant result. Bergmann et al (2000) compared the results of the MWW test for non-normal data provided by 11 different statistical packages. They note that the different packages delivered p values ranging "from significant to non-significant at the 5% level, depending on whether a large-sample approximation or an exact permutation form of the test was used and, in the former case, whether or not a correction for continuity was used and whether or not a correction for ties was made". They concluded that "the only accurate form of the Wilcoxon-Mann-Whitney procedure is one in which the *exact permutation null distribution* is compiled for the actual data".

The MWW test is based on the U statistic where:

$$U = \Sigma_{i=1}^{n_1} \Sigma_{j=1}^{n_2} \phi(x_i, y_j) \tag{1}$$

and

$$\phi(x_i, y_j) = \begin{cases} 1 \text{ if } x_i > y_j \\ 0 \text{ if } x_i \leq y_j. \end{cases} \tag{2}$$

The Wilcoxon test is based on converting the data from two independent datasets G1 and G2 of size n_1 and n_2 respectively into ranks where the ranks are based on all the data (irrespective of which group an observation belongs to). The test statistic (W) is the sum of ranks of observations in G1:

$$W = U + \frac{(n_1 + 1)\, n_1}{2} \tag{3}$$

The statistical tests for W and U are both based on the assumption that there are no duplicate values.

However, ranks have a number of specific properties, that can be seen by considering the formulas for the sum of N integers and the sum of N squared integers:

$$R = \Sigma_i^N i = \frac{(N + 1)\, N}{2} \tag{4}$$

This means the average rank is

$$\overline{R} = \frac{R}{N} = \frac{N + 1}{2} \tag{5}$$

Also

$$\Sigma_i^N i^2 = \frac{(N + 1)\,(2N + 1)\, N}{6} \tag{6}$$

which means that the variance of N ranks is:

$$s_R^2 = \frac{N\,(N + 1)}{6} \tag{7}$$

The equations for the mean and variance of ranks make it clear that, unlike the mean and variance of the raw variables, ranks can never converge to a finite mean and variance. As the number of observations increase, the mean and variance of the ranks increase. Furthermore, if sample sizes are unequal and the null hypothesis is false (i.e., the groups differ), we are almost certain to find large differences in the variances of each group. This variance instability makes applying the large sample tests, which are equivalent to applying the t test (or the F test for multiple groups) to the ranks, very unreliable. This is the reason why the rank transform process proposed by Conover and Imam (1981) is invalid[3]. In addition, the values of U and W depend on the number of observations, so they do not lead to a meaningful effect size.

Looking back to the definition of U, we can see that it is related to the probability that a random observation from one group is larger than a random observation from another group. Other more reliable non-parametric effect sizes are based on normalising U with respect to the sample size and are discussed in Section 3.3.

3 Robust Statistical Methods

Firstly we consider the use of kernel density plots to visualise the distribution of data sets. Then, we present various robust statistical methods described by Wilcox (2012), who also provides **R** algorithms implementing them at his website[4].

3.1 Kernel density plots

In the past, Kitchenham recommended the use of box plots to give researchers an overview of the distribution of a data set, which could alert them to potential problems of non-normality[5]. Now, we believe that advice to be incorrect, and that kernel density plots are often preferable. Kernel density plots are derived from smoothing histograms. Algorithms that construct kernel density plots are available in the **R** language (R Core Team, 2015).

Figure 1 shows a box plot and two histograms with their kernel density plots superimposed. The data set in each case is the same. It is a data set of development effort (man hours) from 38 Finnish projects (Kitchenham and Känsälä, 1983). The box plot and both of the kernel density plots suggest that the data is skewed. The box plot in Figure 1 (a) shows the median slightly off-centred in the box towards the origin, and has a short lower tail and a long upper tail with a single large outlier. However, the kernel density

[3] Using the rank transform process, data are converted to ranks and a standard parametric analysis is applied to the ranked data rather than the raw data.

[4] http://college.usc.edu/labs/rwilcox/home

[5] There are still many circumstances when a box plot can be extremely useful, for example when comparing a large number of related distributions.

plots in Figures 1 (b) and (c) provide more detail concerning the distribution, for example indicating the possibility that the distribution is bi-modal and confirming that the majority of the values are relatively close to the origin. The kernel density plots are different because the number of bins used in each density plot is different, however, the general shape of the two functions is very similar. This demonstrates one major advantage of kernel density plots: they are not so dependent on bin size as histograms.

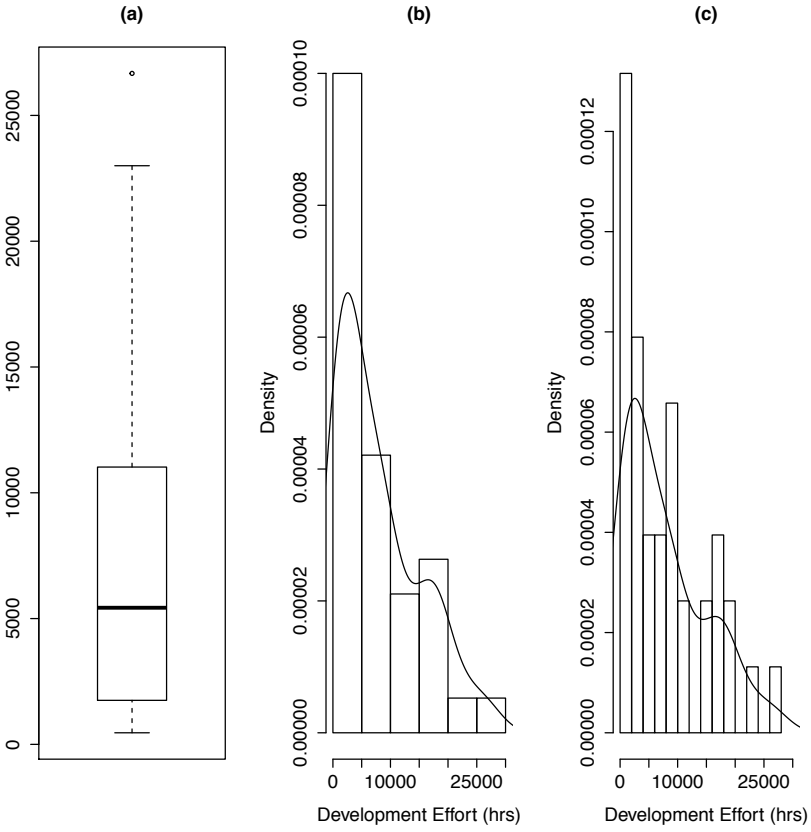

Fig. 1: A Boxplot and two Kernel Density Plots of the same Finnish data set

Figure 2 shows the box plot and kernel density plot of the same 38 projects after transforming the data by taking logs. The box plot suggests that the transformation has improved the distribution of the data. However, the kernel density plot suggests that the direction of skewness has been changed and the data is still far from normal.

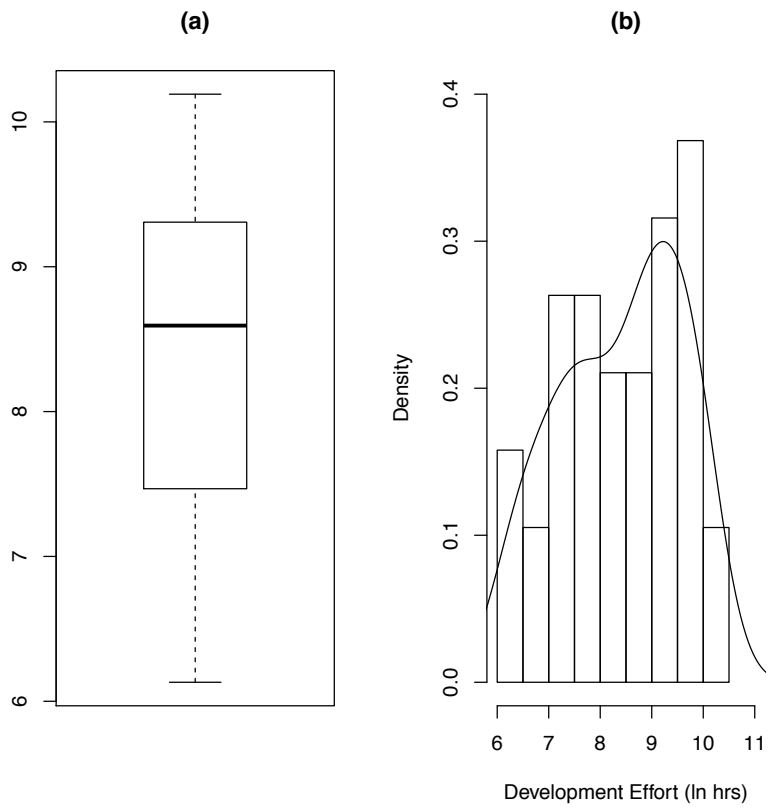

Fig. 2: Box plot and kernel density plots of transformed effort data

Figure 3 presents results from a study of software defect prediction methods aimed at comparing simple product based models with models including product metrics and a process metric (Madeyski and Jureczko, 2015).

It shows the box plots of the percentage of classes that need to be tested to find 80% of the defects using a simple product-based model and an advanced model including a process metric. The data is based on 34 software projects (Madeyski and Jureczko, 2015; Madeyski, 2015). Looking at the box plots of the raw data many of us would believe it was acceptable to use a paired t-test to determine whether the advanced algorithm was better than the simple algorithm (that is, required fewer classes to find 80% of defects). It is not until we view the box plot of the difference between the raw data values in Figure 3 (c) that we see any indication of the problem with this data set.

However, looking at the corresponding kernel density plots in Figure 4, it is clear that distributions of the two sets of observations, Figure 4 (a) and (b), are

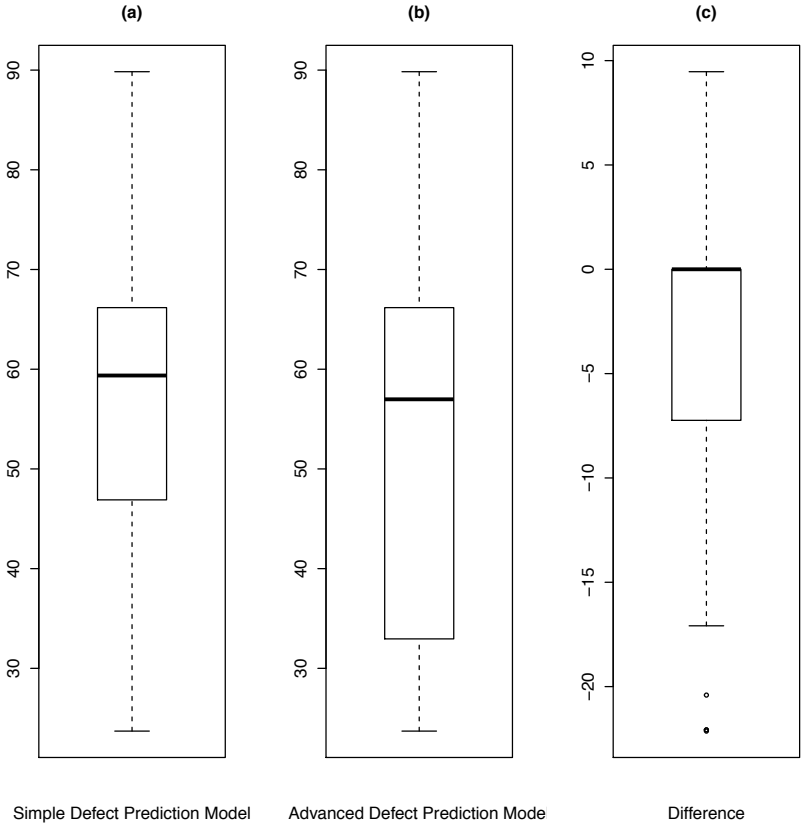

Fig. 3: Box plots of software defect prediction data

far from normal. They have long tails and are possibly bi-modal. Furthermore, the kernel density plot of the difference between the paired outcome values in Figure 4 (c) looks even worse. Although the density close to the origin looks fairly normal, it is clear that the data has a very long lower tail with several extreme values. Looking back to Figure 3 (c), we can see that this is a case where the box plot provides additional useful information. Although the kernel density plot of the difference between the paired observations in Figure 4 (c) seems normal close to the origin, the corresponding box plot indicates that there are many difference observations that share the same zero value, so the distribution is strongly non-normal at the origin.

Overall these examples suggest that the use of kernel density plots and histograms are more likely to alert us to non-normal data than box plots, but box plots can also provide useful additional information.

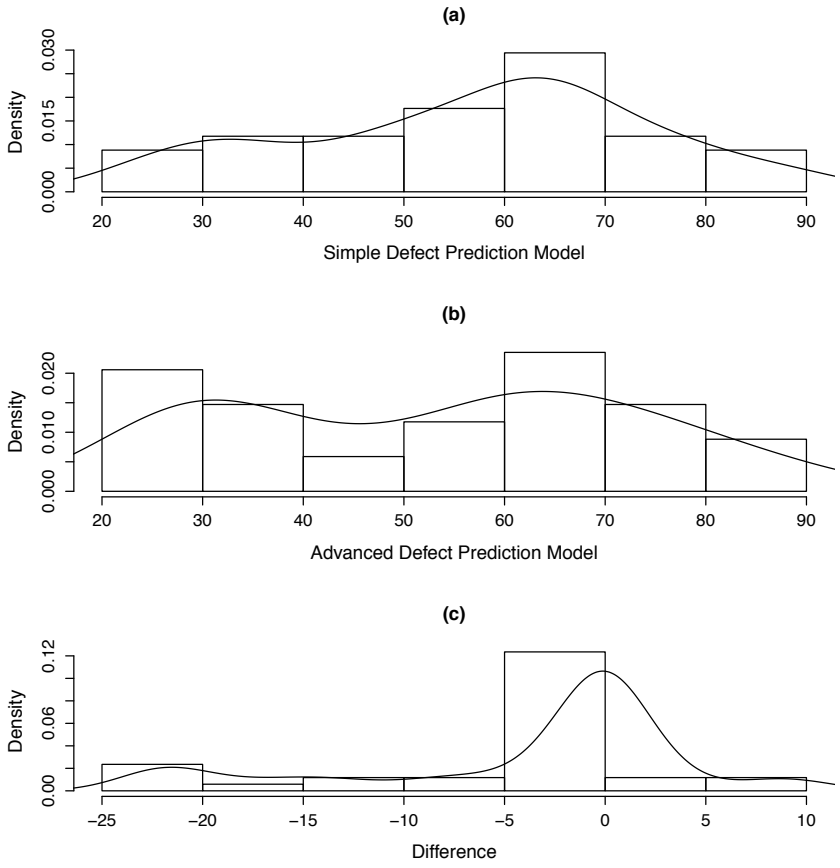

Fig. 4: Kernel density plots of software defect prediction data

3.2 Robust Parametric Methods

One of the most well-known robust metrics of central location is the median. It is, however, not ideal. Although the median is robust, it ignores all but one or two observations. This means that estimates of the standard error of the median are not efficient. They may also be unreliable if there are duplicate values in the data. Price and Bonett (2001) have evaluated several estimators of the sample median and proposed a new estimator that tends to have the smallest bias.

Another common approach is to remove outliers and then use the standard mean and variance of the remaining data. Wilcox and Keselman (2003) point out that there are two problems with this approach:

1. Outlier detection methods based on means and standard deviations can fail to detect outliers.
2. When extreme values are discarded, the remaining observations are no longer independent, which invalidates the calculation of the standard error.

However, Wilcox (2012) introduces several robust measures based on removing outliers through the use of a reliable method of detecting outliers. A related approach is called *trimming*. This means removing a fixed proportion of the smallest and largest values in the data set. These methods are explained below.

3.2.1 Robust Measures Based on Outlier Detection

Robust outlier detection relies on a robust measure of scale such as the median of the absolute deviations from the median (MAD), so if M is the median of a set of n observations:

$$MAD = median|x_i - M|_{i=1,\ldots,n} \tag{8}$$

In the case of data from a normal distribution, MAD estimates the standard deviation multiplied by $z_{0.75} = 0.6745$, which is the 0.75 quantile of the standard normal distribution (that is a distribution with mean $\mu = 0$ and variance $\sigma = 1$). Any observation from the distribution has a 0.5 probability of being within plus or minus 0.6745 of the median. Therefore, instead of MAD, analysts usually use $MADN$, where:

$$MADN = \frac{MAD}{0.6745} \tag{9}$$

$MADN$ is preferred because, if the set of observations is normally distributed, it is an unbiased estimate of the standard deviation. $MADN$ can be therefore be considered a robust measure of *scale*.

A value x_i is then assumed to be an outlier if:

$$\frac{|x_i - M|}{MADN} > k \tag{10}$$

For outlier detection, Wilcox recommends setting k to 2.24. The value 2.24 corresponds to the 0.9875 quantile of the standard normal distribution. This criterion appears less severe than using the theoretical upper and lower tail points of the box plot as a criterion for outlier detection, which corresponds to $z_{0.9965} \approx 2.698$.[6] However, in practice, the upper (lower) tail length of a box plot is decreased because the theoretical value of the upper (lower) tail is shrunk to the nearest actual data value

To construct a robust measure of central location M_{est}, k is set to 1.28:

$$M_{est} = \frac{1.28\,(MADN)\,(i_2 - i_1) + \Sigma_{i=i_1+1}^{n-i_2} x_{(i)}}{n - i_1 - i_2} \tag{11}$$

[6] The theoretical value of the upper (lower) tail of the box plot equivalent is found by multiplying the box length (which calculated as $z_{0.75} - z_{.25}$) by 1.5 and adding (subtracting) it to the upper fourth (from the lower fourth).

where $x_{(1)}, x_{(2)}, ..., x_{(N)}$ are the observations written in ascending order, i_1 is the number of points for which $(x_i - M)/MADN < -1.28$ and i_2 is the number of points for which $(x_i - M)/MADN > 1.28$. The value 1.28 corresponds to the 0.9 quantile of the standard normal distribution, which means that a randomly sampled observation will have an 80% chance of being between plus or minus 1.28. Wilcox notes that this value is often used in the construction of robust estimators because it guards against relatively large standard errors but sacrifices very little data when sampling from a normal distribution.

Initially, $MADN$ is constructed using the median of the raw data. If the estimation process is stopped at that point M_{est} is referred to as the *one-step M−estimator* (MOS). However, M_{est} can be iteratively refined by substituting the current value of M_{est} for the median when calculating $MADN$ in the next iteration. We explain the theoretical justification for M_{est} in Appendix A. Wilcox provides a bootstrap method for calculating the standard error of M_{est}, but this must be treated with caution unless our data set is a random sample from a defined population.

Omitting the term $1.28\,(MADN)\,(i_2 - i_1)$ and replacing the criterion for identifying an outlier with $k = 2.24$, leads to another estimate called the modified one step M-estimator (MOM). Wilcox notes that MOS is better in terms of the size of the standard error, but MOM has advantages when using small sample sizes to test hypotheses. Wilcox provides a bootstrap method for calculating the confidence limits of MOM but does not provide an estimate of the standard error.

3.2.2 Trimmed and Winsorized Means

Trimmed means are based on removing the X% smallest and largest values in a data set. The optimum value of X is unknown but 20% is a reasonable default. Wilcox suggests that this provides a reasonable balance between achieving a small standard error and controlling the probability of a Type 1 error. The observations in the data set that specify the values that correspond to the bottom and top X% of observations are calculated as follows. The data needs to be sorted in ascending and given subscripts from 1 to N identifying that order. Then the subscript of the observation corresponding to $X/100 = 0.0X$ quantile, has the subscript:

$$i_{bottom} = floor(0.0X \times N) + 1 \tag{12}$$

and the subscript of the observation corresponding to $1 - 0.0X$ quantile is

$$i_{top} = N - i_{lowest} + 1 \tag{13}$$

where the function *floor* truncates the value of its parameter to the nearest integer. Then, all observations with values lower than the value corresponding to i_{bottom} and all observations with values greater than i_{top} are excluded from calculation of the trimmed mean.

Winsorized means are derived by replacing the $X\%$ lowest observations with the value of the $X\%$ quantile and $X\%$ largest observations with the value of the $(100 - X)\%$ quantile. This is referred to as *Winsorizing* the data. All observations with subscripts lower than i_{bottom} are replaced by the value of the observation with subscript equal to i_{bottom}. All observations with subscript greater than i_{top} are replaced by the value of the observation with the subscript i_{top}.

Trimmed means form the basis of alternative approaches to t and F. Winsorized means, however, are not usually used as robust central measure in their own right. They are used as a means of obtaining the variance of trimmed means. If a data set of N data points is Winsorized and the estimate of the variance of the Winsorized data set is s_w^2 calculated in the usual way, s_w^2 is another robust measure of *scale*. Furthermore, the estimate of the variance of trimmed mean is:

$$s_{tr}^2 = \frac{s_w^2}{N\,(1 - X/100)} \tag{14}$$

The square root of s_{tr}^2 is the standard error of the trimmed mean.

3.2.3 Examples of Robust Measures of Central Location and Spread

The goal of robust measures of central location and spread is to be resistant to "misbehaviour in the data". We identify the mean as non-robust because one very large abnormal value could make the mean value abnormally large. In contrast, the median is considered robust because one very large abnormal value would not have any effect on the median. This property is shared by all the other robust metrics discussed in Section 3.2.1 and Section 3.2.2 which either remove abnormally large and abnormally small values or replace them. However, unlike the data sets used in our examples, in industry data sets are not static. They grow as new projects are completed and existing products are updated. To investigate the impact of data set growth, we look at how the robust metrics behave when the largest value is removed

In Table 1 we report various measures of central location derived from the data set shown in Figure 1. We report the values from the full data set and from the data set after the maximum effort value was removed.

Considering first the metrics derived from the full data set, we see that, as might be expected in a highly skewed data set, the mean is the largest of the central value metrics and the median is the smallest. The M_{est}, MOM and MOS are all derived in a similar way and all have similar values, in fact M_{est} and MOS have identical values. The mean has the smallest standard error while the standard error of the other metrics (for which standard errors can be calculated) are similar.

Looking at the impact on the metrics after removing the maximum value from the data set, we can see that all the values have been reduced. The median has exhibited the largest percentage change (11%). This might be considered unexpected because the median is supposed to be resistant to changes at the

extremes of the data set. It occurs because the values in the data set consist of only 38 data points, which are spread over a very large range of values (from 460 to 26670). The data points in the centre of the data set are not close together, so when a data point is removed, it causes a large fluctuation in the median. Originally, the median was calculated as the average of the two central values ($5430 = (4830 + 6030)/2$), once the maximum was removed the median became the central value of the remaining 37 values which is 4830.

Of the other metrics, most exhibited a change of between 6% and 7%, including the mean. The mean was not as affected by the removal of the largest value as might be expected because there were a relatively large number of large values in the data set. In this case, the Winsorized mean exhibited the smallest change because with 38, the observation with $i_{top} = 31$ corresponded to an observation with value 14568. Once the maximum value was removed, the value of $i_{top} = 30$ corresponded to an observation with the value 14504, corresponding to a very small 0.4% change in the maximum value of the Winsorized data set. In terms of the effect of removing the maximum value on the standard error, as expected, the standard error of the mean exhibited the largest change, and the standard error of the trimmed mean exhibited the smallest change.

As another example, consider the original COCOMO data set (Boehm, 1981). This contains data on 63 software projects including staff effort measured in person hours and project size measured in K adjusted delivered source instruction (AKDSI)[7], from which we can estimate productivity as $AKDSI/Effort$. The box plot and kernel density plot of the productivity data are shown in Figure 5. Both the box plot Figure 5 (a) and the kernel density plot Figure 5 (b) agree that the data is highly skewed and contains outliers. In contrast to the Effort data set, this data set is concentrated over a small range (0.020465 to 1.25). In addition, the largest value is relatively far from the next smallest value 0.8833), and the central five values are very close together $(0.18408, 0.1917, 0.1923, 0.1987, 0.2000)$. The robust measures for the productivity data are shown in Table 2.

Given the properties of this data set it is not surprising to find that the mean exhibits a large change when the maximum value is removed and the median exhibits only a small change. In this case, the Winsorized mean exhibits the largest change. This is because with the full data set, $N = 63$, and the value of i_{top} was 51 corresponding to an observation with the value 0.4333. Once the maximum value was removed, the value of i_{top} was 50 corresponding to an observation with the value 0.3786. This corresponded to a relatively large 12.6% change in the maximum value of the Winsorized data set. In this case, most of the standard errors exhibited a relatively large change with the change to the median standard error being the largest (27.0%).

These examples, might suggest that resistance is a somewhat relative concept in the context of evolving data sets and depends on the specific nature

[7] The adjustment occurs when projects are updated rather than created as new, and is intended to reflect the amount of new/changed lines of code needed to produce the update.

of a data set. However, they confirm that for skewed data with outliers, the trimmed mean will be closer to the central point of the data set than the mean and will usually be smaller than the $M - Estimator$, MOS or MOM. It will also usually have a smaller standard error than the mean, even though the divisor (and associated degrees of freedom) will be based on $N(1 - 0.0X)$ rather than N.

However, the real importance of using trimmed means and other robust parametric measures is that they allow non-normal data to be analysed fairly on the raw data scale. This is particularly important for ratio-based measures that are known to be strongly skewed, such as productivity (effort/size) or defect rates (faults/size). In spite of the extreme non-normality of such data, practitioners still prefer to use average productivity metrics based on the raw data, for example, to set up baselines and identify good practice, see for example Huijgens et al (2013).

The problem with using the mean is that with skewed data more than 50% of projects have productivity values less than the mean. In the COCOMO productivity data, 62% of the projects had productivity values less than the mean productivity value. Using the mean value gives an inflated value to the central location of the data set, as a result of the large values. The median is much smaller than the mean and 49% of the projects are less than the median. However, since the median is only based on one or two values (depending on whether the data set has an odd or even number of observations), it is hard to defend the median as a trustworthy measure. In contrast to the mean, 54% projects had productivity values less than the trimmed mean. Furthermore, since the trimmed mean is based on 60% of the data set it is a more defensible estimate of the central location than the median.

The practical implication is that benchmarking initiatives that label projects with values less than the mean as *poorly-performing* projects might justifiably be rejected by project managers whose projects performed better than the median. In the case of the COCOMO productivity data, five projects had values greater than the trimmed mean but less than the mean. Furthermore, if the data did not include the largest value, none of the projects would change from being classified as above the trimmed mean to below the trimmed mean.

We would also suggest that projects within *plus or minus* two standard errors of the trimmed mean should be considered as exhibiting *average* productivity. Using this criterion, the trimmed mean would classify projects with order statistics $i = 28$ to $i = 39$ as being average, and there would be no change if the largest value were removed. In contrast, using the mean and its standard error, the nine projects with order statistics $i = 35$ to $i = 43$ would be classified as average, and if the largest value were removed, the mean would classify the 8 projects with order statistics $i = 36$ to $i = 43$ as being average. Bearing in mind that the median value corresponds to the project with order statistic $i = 32$, it is clear that using the trimmed mean identifies more projects close to the centre of the distribution as average than does the mean.

To identify poorly and exceptionally performing projects, observations with productivity values less than the value of the observation corresponding to i_{bottom} could be described as poorly performing (in the COCOMO example, the observation with $i = 13$ which had a value 0.07266 corresponded i_{bottom}). Equally, projects with productivity values greater than the value of the observation corresponding to i_{top} could be described as exceptionally performing projects (in the COCOMO example the observation with $i = 51$ which had a value 0.4333 corresponded to i_{top}). Huijgens et al (2013) point out the value of investigating whether poorly performing projects and exceptionally performing projects have specific characteristics. In the case of the COCOMO productivity data, all of the poorly performing projects were categorized as *embedded projects*, while the projects with the six largest productivity values were all classified as *organic projects* and the remaining six exceptionally performing products were classified as *semi-detached projects*. In the next section, we follow up the issue of the impact of project type on productivity in order to demonstrate how trimming can be used to test hypotheses about non-normal data sets on the raw data scale.

Table 1: Central Location and Scale measures for the Effort Data with and without maximum value

Metric Name	Central Location	Standard Error	Central Location without maximum (%age Change)	Standard Error without maximum (%age Change)
Mean	7678.2895	1157.4953	7165 (6.68%)	1065.8918 (7.91%)
Median	5430	1522.0595	4830 (11.05%)	1626.3678 (6.85%)
M-Estimator	6634.2307	1560.7222	6206.4239 (6.45%)	1484.903 (4.86%)
MOS	6634.2307	NA	6206.4239 (6.45%)	NA
MOM	6377.2857	NA	5658.697 (11.27%)	NA
20% Trimmed Mean	6123.4583	1414.9294	5756.3043 (6%)	1403.2146 (0.83%)
20% Winsorized Mean	6796.0263	1365.7145	6573.8649 (3.27%)	1377.7016 (0.88%)

Table 2: Central Location and Spread of Productivity data with and without the maximum value

Metric Name	Central Location	Standard Error	Central Location without maximum (%age Change)	Standard Error without maximum (%age Change)
Mean	0.2725	0.0316	0.2568 (5.78%)	0.0278 (11.96%)
Median	0.1923	0.0387	0.192 (0.17%)	0.0283 (26.98%)
M-Estimator	0.2251	0.0313	0.2206 (2.03%)	0.0298 (4.83%)
MOS	0.2256	NA	0.2209 (2.08%)	NA
MOM	0.203	NA	0.203 (0%)	NA
20% Trimmed Mean	0.2092	0.0291	0.2033 (2.82%)	0.0256 (11.74%)
20% Winsorized Mean	0.2259	0.0284	0.212 (6.17%)	0.0254 (10.81%)

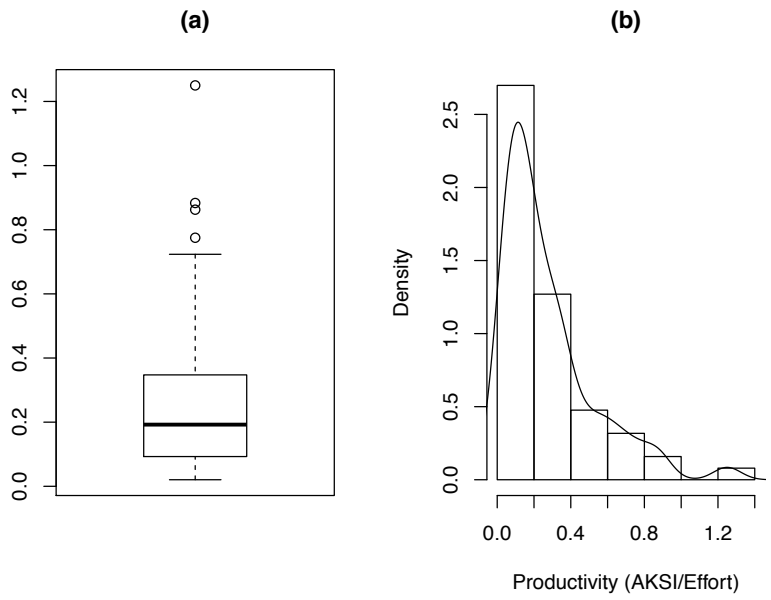

Fig. 5: COCOMO Productivity Data — All Projects

Another important issue is that robust measures of spread can be generalised into robust measures of covariance. This leads to the ability to undertake multivariate analysis and robust regression analysis of non-normal data sets without relying on normalising transformations. Although it is beyond the scope of this paper, Wilcox (2012) discusses multivariate methods and robust regression extensively.

3.2.4 Robust alternatives to t and F tests

The problem associated with heteroscedasticity among different samples has been known for a long time. Welch (1938) proposed a variant of the t-test that allowed for different variances within each group. This is the *default* version of the t–test in **R** (R Core Team, 2015).

The variance of the difference between two means is calculated as:[8]

$$var\left(\bar{x}_1 - \bar{x}_2\right) = \frac{\left(n_1 - 1\right) s_1^2 + \left(n_2 - 1\right) s_2^2}{\left(n_1 + n_2 - 2\right)} \tag{15}$$

where $\bar{x}_1$ is the mean of the n_1 data points in one of the groups and $\bar{x}_2$ is the mean of the n_1 data points in the other group. This is very similar to the

[8] This equation and the equation for the degrees of freedom are incorrect in (Kitchenham, 2015).

original t-test except that the two variances are not combined into an overall average. The major difference between a Welch test and a t-test is that the degrees of freedom are calculated quite differently as:

$$df = \frac{(q_1 + q_2)^2}{\left(\frac{{q_1}^2}{(n_1-1)} + \frac{{q_2}^2}{(n_2-1)}\right)} \quad (16)$$

where $q_i = s_i^2/n_i$.

Yuen's test uses trimmed means instead of the ordinary means together with Welch's test as a robust test for comparing the central location of two sets of data (Yuen, 1974). This approach can be extended to cater for repeated measures (paired) designs, multiple groups and factorial designs. It also allows researchers to test linear combinations among mean values. For example in a factorial experiment a researcher might want to know if three levels of a factor are additive. For example, suppose we have a cost estimation factor such as "Required reliability" that has three levels "Low", "Standard" and "High", and we believe that this has an additive effect on productivity. If we have productivity values for projects with the different levels of reliability, an additive hypothesis is tested using the following linear combination of mean values:

$$\hat{x}_{Standard} - \hat{x}_{Low} = \hat{x}_{High} - \hat{x}_{Standard}$$

or equivalently, that

$$\hat{x}_{Low} + \hat{x}_{High} - 2\hat{x}_{Standard} = 0$$

Yuen's method is appropriate when testing for differences between central locations, but would not be sensitive to changes in the lower tail of a distribution of the kind that can be seen in Figure 4.

A disadvantage of the use of Yuen's method is that the use of trimming and Welch's test means that the number of degrees of freedom are substantially reduced. This will mean we need more observations. However, if our data are not normal, we will also need a great many observations before we can be sure that results based on the full data set are reliable.

As an example of this approach, consider the original COCOMO data set (Boehm, 1981). As discussed in 3.2.3, the projects were divided into three different types (referred to as the project *mode*), labelled *organic*, *embedded* and *semi-detached*. Using this data it is possible to test whether the productivity of projects of each type is the same.

The histograms and kernel density plots for projects of each type are shown in Figure 6. Inspection of the plots confirms long tailed distributions. It also suggests that productivity is generally highest for organic projects and lowest for embedded projects, with semi-detached projects somewhere in between.[9]

[9] Comparing Figure 5 and Figure 6, also confirms that analysing data sets in more homogeneous subsets is likely to make the distribution of the subsets less pathological than the distribution of the full data set.

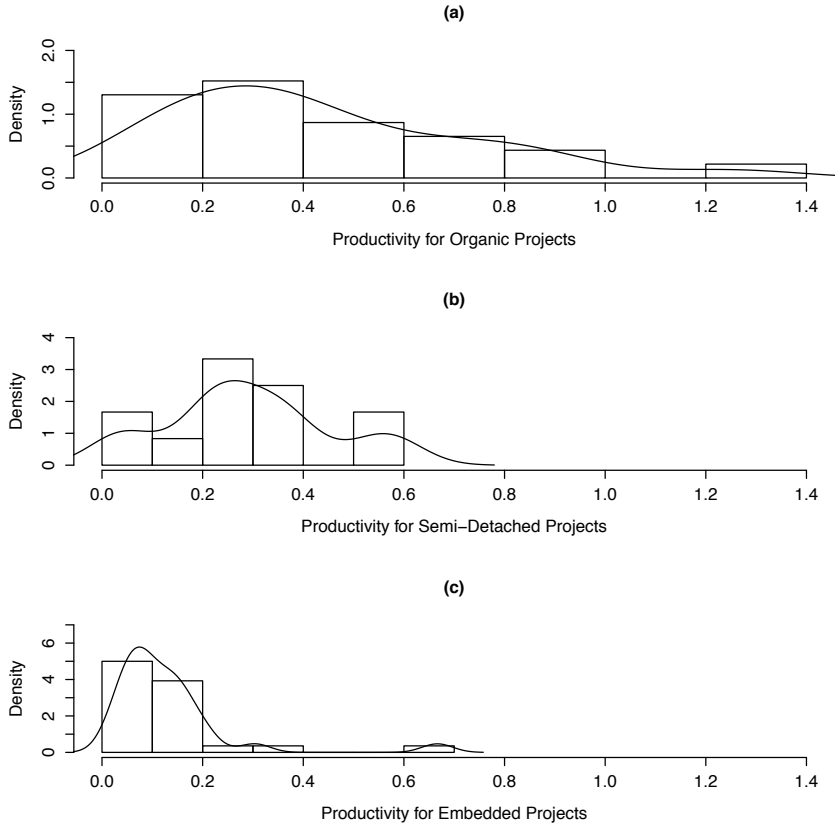

Fig. 6: Kernel density plots for the COCOMO data set for each project type

Using trimmed means and the algorithms produced by Wilcox, we can test whether there are significant differences among the trimmed means for the different modes and also whether there is a linear relationship between trimmed means (Wilcox, 2012). Summary statistics of the COCOMO project productivity values are shown in Table 3.

Table 3: COCOMO Project Productivity Summary Statistics

Project Type	# Projects	Mean	SE	Trimmed Mean	TM SE
Organic	23	0.4368	0.0625	0.3901	0.0718
Semi-detached	12	0.291	0.0482	0.285	0.0375
Embedded	28	0.1296	0.0233	0.1052	0.0133

Using Yuen's method, an overall F-test for differences among the three groups of projects was statistically significant ($F = 18.678$, $df_1 = 2, df_2 = 14.74, p = 9.100371e-05$). Although there are 28 embedded projects, 12 semi-detached projects and 23 organic projects, the degrees of freedom for the denominator of the F-test is 14.74 rather than the 60 that would be found in a standard analysis of variance. This is because 40% of the data is removed by trimming and the use of Welch's method for unstable variances further reduces the degrees of freedom and results in non-integer values for degrees of freedom.

Wilcox also provides an algorithm that assesses all pairwise comparisons of the trimmed means that also adjusts the confidence intervals to allow for multiple tests. The results of this analysis are shown in Table 4. This suggests that both organic and semi-detached projects are more productive than embedded projects but that there is no significant difference between semi-detached and organic project productivity.

Table 4: COCOMO Project Productivity Group Comparisons

Comparison	TM Difference	Lower 95% CL	Upper 95% CL	df
E v SD	−0.1798	−0.2863	−0.0733	8.8933
E v O	−0.2849	−0.4664	−0.1034	14.9841
SD v O	−0.1051	−0.3004	0.0902	19.6753

Wilcox's algorithm will also allow you to test linear combinations of the trimmed means, for example, to test the hypothesis that the difference between the trimmed means is linear, that is:

$$LinearCombination = TM_E + TM_O - 2TM_{SD} \approx 0 \tag{17}$$

The value of the linear combination of trimmed means for the COCOMO data is −0.7706 with 95% confidence limits (−0.2779, 0.1284). This indicates that we cannot rule out the possibility of a linear effect. However, the degrees of freedom for this test is 18.85, which suggests the test has a low power, which is particularly problematic if we want to be confident that the null hypothesis is likely to be true.

3.3 Non-parametric tests

Looking at Figure 4 rather than considering the difference between means, it might be useful to ask the question "What is the probability that a random observation from the set of simple algorithms is greater than an observation from the set of advanced algorithms". This question is the rationale behind Cliff's δ. It is also similar to the rationale for the MWW test with the U

statistic, but unlike U, δ can cope with duplicate values. First of all we need to consider three probabilities:

$$p_1 = P\left(x_{1i} > x_{2i}\right)$$

$$p_2 = P\left(x_{1i} = x_{2i}\right)$$

$$p_3 = P\left(x_{1i} < x_{2i}\right)$$

Then Cliff's δ is defined as:

$$\delta = p_1 - p_3 \tag{18}$$

and is therefore the difference between the probability that a random observation from group one is greater than a random observation from group two and the probability that a random observation from group one is less than a random observation from group two (Cliff, 1993). This is also called the expanded success rate difference (SRD) (Kraemer and Kupfer, 2006).

p_1 and p_2 can be used to calculate the probability of superiority (Grissom, 1996):

$$\hat{P} = p_1 + 0.5p_2 \tag{19}$$

This metric has also been called the area under the receiver curve (AUC) (Kraemer and Kupfer, 2006), the measure of stochastic superiority ($\hat{A_{12}}$) (Vargha and Delaney, 2000) and the probabilistic index ($P(X > Y)$) (Acion et al, 2006). Arcuri and Briand (2014) recommend using the metric for software engineering data analysis. Following Vargha and Delaney (2000), they refer to it as the $\hat{A_{12}}$ metric. Also since $p_1 + p_2 + p_3 = 1$:

$$p_3 = 1 - \hat{P} - 0.5p_2 \tag{20}$$

which means

$$\delta = 2\hat{P} - 1 \tag{21}$$

Cliff derived the standard deviation for δ, which can be used to calculate the standard deviation of $\hat{P}$, since $var_\delta = 4var_{\hat{P}}$.

Looking at the calculation of the MWW U shown in Equation 1, and assuming there are no duplicate values:

$$p_1 = \frac{U}{n_1 n_2} \tag{22}$$

and

$$p_3 = 1 - \frac{U}{n_1 n_2} \tag{23}$$

so, that

$$\delta = \frac{2U}{n_1 n_2} - 1 \tag{24}$$

Akritas and Arnold (1994) and Brunner et al (2002) suggested a different but related method, which also allows for duplicate observations by using midranks. Midranks are necessary if there are two (or more) observations with the same value, in that case, the observations are both allocated the average of

the two (or more) related ranks. Their method is an ANOVA-like method based on ranks but is robust to heteroscedasticity of group variances. It is important because it can be used to analyse much more complicated statistical designs than simple between-groups designs.

In the two group case, their test metric is simply the probability of superiority, $\hat{P}$. It is calculated by first pooling all observations and calculating all R_{ij} which are the midranks associated with the observations x_{ij} where ij corresponds to the ith observation in group j. The average rank for group j is:

$$\hat{R}_j = \frac{1}{n_j} \Sigma_{i=1}^{n_j} R_{ij} \tag{25}$$

Then:

$$\hat{P} = \frac{1}{n_1 + n_2} \left(\bar{R}_2 - \bar{R}_1 \right) + 0.5 \tag{26}$$

Wilcox (2012) reports that there is not much to choose between Cliff's method and Brunner et al.'s method, but that Cliff's method may have some advantages when there are many tied values and sample sizes are small.

The ANOVA-like method can be generalised to multiple groups, but to perform an overall test of differences among $k > 2$ groups uses *relative effects*, where the relative effect for group j is calculated as:

$$\hat{q}_j = \frac{\left(\hat{R}_j - 0.5 \right)}{N} \tag{27}$$

where $N = \Sigma n_j$. The null hypothesis in this case is that $q_1 = q_2 = ... = q_k = 0.5$.

Both the ANOVA-like method and Cliff's method can be adapted for repeated measures. This means they could be considered for analysing the software defect prediction data shown in Figure 4. However, since the data includes 17 duplicates, it is likely that Cliff's δ is more appropriate. Wilcox does not provide an implementation of the paired data test, but the details are provided in (Cliff, 1993). The estimate of δ is[10]:

$$d_w = \frac{\Sigma d_{ii}}{n} \tag{28}$$

where $d_{ii} = 1$ if the simple software defect prediction model identified fewer classes than the advanced model, 0 if the models identified the same number of classes, and -1 if the simple model identified more classes. The variance of d_w is:

$$s_{d_w}^2 = \Sigma \frac{(d_{ii} - d_w)^2}{n(n-1)} \tag{29}$$

The estimated value of Cliff's δ is -0.2647 with 95% confidence interval $(-0.4884, -0.0410)$ on the assumption that the estimate is approximately normally distributed. The test value is -2.319 which has a probability of

[10] We use d rather than δ when referring to sample-based estimates of δ.

$p = 0.0102$. This suggests that the predictions made by the advanced defect prediction algorithm have a significant probability of requiring the search of fewer classes than the simple algorithm. This can be compared with the standard Wilcoxon test which reports a $p-$value of 0.01577 but delivers a warning "cannot compute exact p-values with zeroes".

For analysing multiple repeated measures (for example, studies where many different cost estimation algorithms are applied to many different data sets), software engineering researchers have often adopted Friedman's test with corresponding post-hoc tests as recommended by Demšar (2006) (see, for example, Dejaeger et al (2012)). However, in a study of the performance of Friedman's test, Agresti and Pendergast (1986) found that for an underlying normal distribution, their rank transformed ANOVA test could be substantially more powerful than the Friedman test. In a more recent paper, Tian and Wilcox (2007) compared the Agresti-Pendergast method with the ANOVA-like method developed by Brunner and colleagues. They found that under most conditions, the ANOVA-like method was preferable to the Agresti-Pendergast method in terms of both Type I errors and power. The exception occurred when there were only two repeated measures for each data set. There has been no direct comparison of the Agresti-Pendergast and Cliff's method for cases where there are only two repeated measures.

3.4 Guidelines for interpreting effect size magnitude

Effect size is a name given to indicators that measure the magnitude of a treatment effect. We agree with Arcuri and Briand (2014) that effect sizes are extremely useful, as they provide an objective measure of the importance of the experimental effect, regardless of the statistical significance of the test statistic. Furthermore, effect sizes are much less affected by sample size than statistical significance and, as a result, are better indicators of practical significance (Madeyski, 2010; Urdan, 2005; Stout and Ruble, 1995).

Cohen (1988, 1992) was the first person to propose interpretation guidelines for effect sizes, by suggesting criteria to define a small, a medium or a large effect for use in the behavioural sciences. However, Cohen did not present any systematic calculation of effect sizes from research studies as the basis for his generalizations. That is why Lipsey and Wilson (2001) found these guidelines somewhat arbitrary, and presented different interpretations of the magnitude of effect sizes based on the distribution of effect sizes for over 300 meta-analyses of psychological, behavioural, and education studies, suggesting the need for domain specific guidelines.

To allow an interpretation of effect sizes in a software engineering context, Kampenes et al (2007) therefore proposed magnitude labels based on a systematic review of effect size in 92 software engineering controlled experiments. The sample size is limited but gives a rough estimation of what constitutes *small*, *medium* and *large* effect sizes in the software engineering domain.

All these guidelines were presented in (Madeyski, 2010). In this paper, the guidelines were extended to include the newest effect size indicators (Cliff's delta and the probability of superiority) and these are all summarised in Table 5.

Table 5: Guidelines for effect size magnitude interpretation

Effect	small	medium	large
(Cohen, 1988)			
d	0.20	0.50	0.80
r	0.10	0.243	0.371
r^2	0.01	0.059	0.138
(Cohen, 1992)			
d	0.20	0.50	0.80
r	0.10	0.30	0.50
r^2	0.01	0.09	0.25
(Lipsey and Wilson, 2001)			
d	0.30	0.50	0.67
(Kampenes et al, 2007)			
g	0.17 $[0.00 - 0.376]$	0.60 $[0.378 - 1.000]$	1.40 $[1.002 - 3.40]$
r	0.09 $[0 - 0.193]$	0.30 $[0.193 - 0.456]$	0.60 $[0.456 - 0.868]$
r^2	0.008 $[0 - 0.0372]$	0.09 $[0.0372 - 0.208]$	0.36 $[0.208 - 0.753]$
Vargha and Delaney (2000); Kraemer and Kupfer (2006)			
Cliffs δ (SRD)	0.112	0.276	0.428
$PS(\hat{A_{12}})$	0.556	0.638	0.714

An important issue for the use of effect sizes in meta-analysis is that the variance of the effect size needs to be estimated. Effect size variances are often quite complex to calculate, but Wilcox's software provides standard errors for the Cliff's d and the probability of superiority (Wilcox, 2012).

4 Example derived from a Multi-site Experiment

This section presents a large-scale example of an analysis using robust methods. In this section, we will demonstrate three different options for analysing our data. However, this is for explanatory purposes only, we do not advocate trying many methods until finding one that gives the answer you want. We return to this issue when discussing the results of the experimental analysis.

4.1 Background to the Multi-site Experiment

The study described here was designed to investigate the use of multi-site studies in order to address the problems of small sample sizes in Software Engineering experiments, see Dybå et al (2006) and Kampenes et al (2007).

The topic for the multi-site experiment concerned the extent to which structured abstracts were clearer and more complete than conventional abstracts. Specifically, the study investigated the following research question:

> Are software engineering researchers likely to produce clearer and more complete abstracts when these are written using a structured form?

A report on our experiences regarding the *organisation* of the multi-site experiment (referred to using the alternative term *distributed experiment*) is provided elsewhere (Budgen et al, 2013). In this paper we are only concerned with the analysis of the data that was collected from this and used to assess the above research question.

4.2 Experimental Design

Formally, our experiment set out to test the following hypotheses:

- *Null Hypothesis 1*: Structured and conventional abstracts written by software engineering researchers are not significantly different with respect to completeness.
- *Alternative Hypothesis 1*: Software engineering researchers write structured abstracts that are significantly more complete than conventional abstracts.
- *Null Hypothesis 2*: Structured and conventional abstracts written by software engineering researchers are not significantly different with regard to clarity.
- *Alternative Hypothesis 2*: Software engineering researchers write structured abstracts that are significantly clearer than conventional abstracts.

To address these, we asked participants to assess the clarity and completeness of abstracts of scientific papers with an empirical element that were published by a Software Engineering journal that had adopted structured abstracts, comparing them with the clarity and completeness of both abstracts published by the same journal before it adopted structured abstracts, as well as with the abstracts published by a similar journal that did not adopt structured abstracts. This gave us the opportunity to see whether the advantages of structured abstracts we had observed in controlled experiments carried over into the field.

4.2.1 Structure and organisation of the multi-site experiment

The abstracts were obtained from academic papers published in the *Information and Software Technology* journal (IST) and the *Journal of Systems and Software* (JSS). These software engineering journals are both published by Elsevier, and contain many papers with an empirical content. The important point for our experiment was that IST began mandating the use of structured abstracts in the time period 2009-2011 whereas JSS retained the use of conventional abstracts.

This experiment is a *quasi-experiment* because we selected abstracts from particular volumes of the two journals, and could not randomise the source of the structured abstracts. Based on the categories provided by Shadish et al (2002), the experiment can be classified as "*a two-group pretest-posttest design with non-equivalent control groups*". Here the change between pretest and posttest is provided by the transition to the use of structured abstracts over the period 2009-2011 for IST, and the non-equivalent control group is provided by the two blocks of abstracts from JSS.

We conducted the experiment across five sites: Durham and Keele Universities (UK), Lincoln University (New Zealand), the City University (Hong Kong), and the Prince of Songkla University (Thailand). Subsequent to the initial experiment two further sets of data were collected, one from students at City University (Hong Kong) and the other from students at Wroclaw University of Science and Technology (Poland). The experiment was organized by Budgen who prepared the experiment protocol and the experimental materials, circulated the relevant materials to each site, and co-ordinated the responses.

An Entity-Relationship style diagram illustrating the experiment together with an explanation of the entities and their relationships is presented in Appendix B.

4.2.2 Independent and dependent variables

For this study we can identify three independent variables:

1. The *source* of the abstracts (JSS; IST)
2. The *time of publication* (Block1; Block2) For both journals, these blocks consist of roughly eighteen months-worth of issues within the period 2009-2011. For JSS, the boundary between blocks was based upon date (mid-2010), whereas for IST, where the transition from conventional to structured abstracts was gradual, with many issues having mixed forms, the boundary is across all issues of 2010, with assignment to block being determined by the form of the abstract.
3. The *location* of the study/participants (UK-2 sites, NZ, Thailand, HK, Poland)

The dependent variables for the study were measures of *completeness* (how well an abstract would enable a systematic reviewer to determine the relevance of the associated paper) and *clarity* (the quality of writing used). These were respectively assessed using a set of 8 questions similar to those employed in previous studies, (Budgen et al, 2011) and (Budgen et al, 2008), and a 10-point Likert-like scale. The completeness score for a specific abstract for each judge was calculated as:

$$Completeness_i = \frac{\Sigma_{i=1}^{i=8}(x_i)}{QA} \tag{30}$$

where x_i is a numeric value for completeness question i where $Yes = 1, No = 0, Partly = 0.5, NA = 0$, and QA is a count of the number of questions

answered, excluding NA responses. Thus the completeness score for a specific abstract by a specific judge is a value between 0 and 1. The completeness score for an abstract is:

$$AbstractCompleteness = median\,(Completeness_i) \tag{31}$$

where $i = 1, ..., 4$.

4.2.3 Participants and their roles

The participants who acted as "judges" of the abstracts were intended to be undergraduate students studying computing in some form, and who were at approximately the same level of technical educational attainment, approximating to two years of specialist computing study at university, but in practice, some universities also recruited participants who were more experienced, see (Budgen et al, 2013). These were students who might be expected to read research papers that have abstracts, but who had not yet had to write dissertations and similar documents containing abstracts. Within the English context (Durham and Keele) this would equate to students who were at the end of their second year of study, or beginning their third year of study. For each site, sixteen participants were recruited locally, using local expertise to match them to the above description. Where necessary, we paid a small honorarium to those taking part. Participants were expected to have a reasonable level of English, since the abstracts were in English, and so we collected data about whether or not this was their first language. Figure 16 is a flow diagram showing a high-level overview of the experimental process undertaken at each site.

Participants were required to act as judges for four abstracts, one taken from IST and one from JSS in the time period prior to the introduction of structured abstracts and one from IST and one from JSS in the time period following the adoption of structured abstracts by IST. In addition, each abstract was evaluated by four judges. A flow diagram of the experimental process from the viewpoint of the judges is shown in Figure 17.

4.2.4 Experimental materials

Budgen identified all empirical papers in JSS and IST over the two defined time periods. The number of abstracts available from each source and each time period is shown in Table 6.

Budgen then created a set of four random number sequences, based on the size of each of the blocks of abstracts. The first four values from each sequence were used to select the abstracts for the first site, the next four for the second site and so on until he had selected 20 abstracts from each journal and each block. In the second data collection activity (from the universities in Hong Kong and Poland), four further abstracts were selected from each journal and block.

Table 6: Allocation of abstracts to blocks

Id	IST organisation	No.	JSS organisation	No.
Block 1	All 2009; conventional (2010)	110	All 2009; Jan-June (2010)	132
Block 2	Structured (2010); all 2011	131	Jul-Dec (2010); all 2011	173
	All IST	241	All JSS	305

All data were collected using paper forms. Budgen prepared a set of data collection forms organized as two A5 sized pages side by side. Each of these had the abstract printed on the right hand page, and the questions on the left hand page. They were also suitably coded so that they could be tracked by the experimenter. To avoid participants guessing which abstract was supposed to be best, Budgen removed the headings from the structured abstracts and revised any sentences rendered ungrammatical by the removal of the headings. In addition, the title and keywords were removed from each abstract.

The questions were derived from those used in the previous studies (Budgen et al, 2011, 2008), with modifications to address the restriction of using only those papers that had an empirical element. For the purpose of data collection, each student judge was required to first complete a consent form, then a short form asking for demographic information, and would then receive the four data collection forms in the defined order[11], and one at a time. As they completed a form it was to be returned to the experimenter, who would check that it had been fully completed and then issue the next form. A flow diagram of the process is shown in Figure 15.

The details of the conduct of the experiment, and of the divergences from the plan that occurred, are described in Budgen et al (2013). The second data collection exercise used the same set of 16 abstracts at two different universities: one in Hong Kong (the City University) the other in Poland (Wroclaw University of Science and Technology).

4.3 Data Analysis

The statistical design is a two-by-two factor analysis with journal as one factor and time period as the other factor. What we are interested in is whether the the difference between completeness of IST and JSS abstracts in the second time period is significantly greater than the difference between completeness of IST abstracts and JSS abstracts in the first time period. This is the interaction term in a factorial design and is sometimes referred to as a *differences in differences analysis*. That is, if we had normally distributed data, we would test whether:

$$\bar{x}_{22} - \bar{x}_{12} - (\bar{x}_{21} - \bar{x}_{11}) > 0 \tag{32}$$

[11] The order was changed for each group of judges that assessed the same abstract

where $\bar{x}_{22}$ is the mean of IST abstracts in period 2, $\bar{x}_{12}$ is the mean of JSS abstracts in period 2, $\bar{x}_{21}$ is the mean of IST abstracts in period 1 and , $\bar{x}_{11}$ is the mean of JSS abstracts in period 1.

This section examines a number of approaches to analysing the data from the experiment using robust methods.

4.3.1 Preliminary Analysis

Data from each site was analysed to assess whether or not there was consensus in the assessment of the abstract completeness score. The analysis was based on a oneway analysis of variance of all abstract data collected at a specific site with "abstract" as the factor with 16 levels. We used a standard ANOVA rather than a robust equivalent for this analysis because we wanted deviations from the mean to be emphasized and to calculate the *Intra-Class Correlation* (ICC) (Shrout and Fleiss, 1979). ICC is assessed on the same subjective scale as the Kappa agreement statistic. The analysis is shown in Table 7 where

$$ICC = \frac{MS_{Between} - MS_{Within}}{MS_{Between}} \tag{33}$$

In Table 7, the column labelled *Phase* identifies whether the data collection took place in the first phase of the experiment or the second; $MSBA$ is the mean squares between abstracts and $MSWA$ is the mean squares within abstracts. It is noticeable that all the sites where English is the first language (that is, Keele University (K), Durham University (D) and Lincoln University (L)) achieved substantial agreement, whereas other sites did not achieve such good agreement, although only the Hong Kong City University data in phase 2 achieved no consensus. Given the lack of consensus, we decided to omit the Hong Kong City University phase 2 data from our subsequent analyses but to include the Hong Kong City University phase 1 data.

Table 7: Agreement among Judges for each site

Phase	Site	$MSBA$	$MSWA$	F	p	ICC
1	Keele	0.0802	0.019	4.2217	0.0001	0.7631($Substantial$)
1	Durham	0.0673	0.0253	2.6598	0.0052	0.624($Substantial$)
1	Lincoln	0.0858	0.0227	3.7767	0.0002	0.7352($Substantial$)
1	Pr. Songkla	0.0409	0.0205	1.9931	0.0362	0.4983($Moderate$)
1	Hong Kong (CU)	0.0463	0.032	1.4494	0.1636	0.31($Fair$)
2	Hong Kong (CU)	0.0429	0.0579	0.7404	0.7322	-0.3506($Poor$)
2	Wroclaw (POLAND)	0.0424	0.025	1.6955	0.0841	0.4102($Moderate$)

4.3.2 Analysis of the Experimental data

The main analysis is in two phases relating to the two data collection periods. In the first phase we analyzed the data from the first 5 sites, in the second phase we used meta-analysis to aggregate the data from both phases.

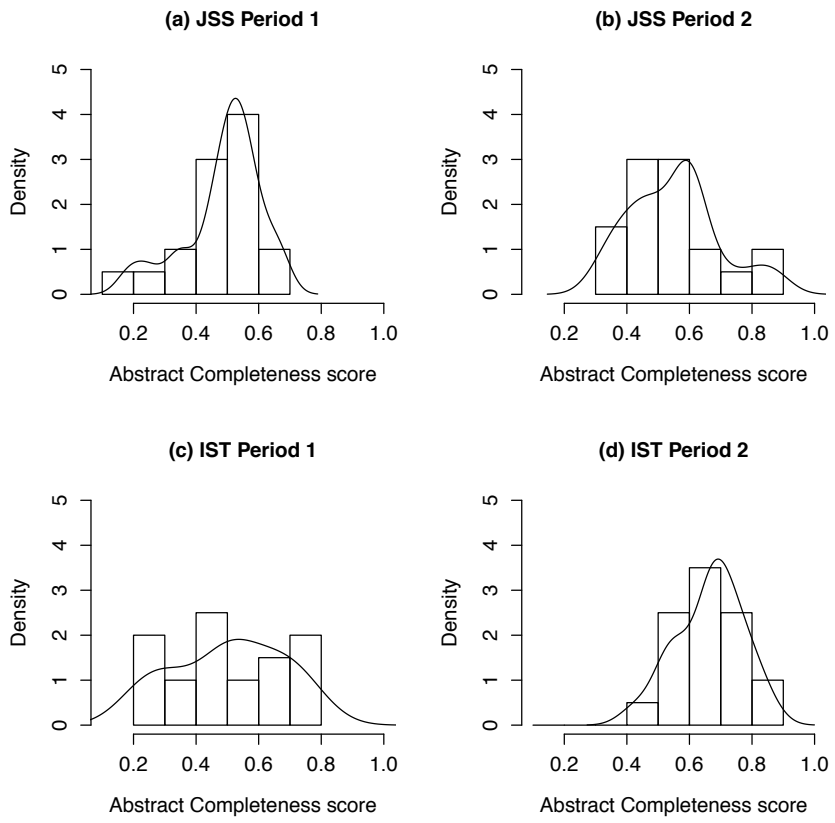

Fig. 7: Analysis of the median completeness score for each abstract

Phase 1 Analysis

Figure 7 shows the kernel density plot of the abstract data from the original 5-sites. This is based on the median of the four average completeness scores for each abstract with 20 abstracts per journal/time period group. We use the median since it is more robust than the mean.

Comparing Figure 7 (d) with plots (a), (b) and (c), it seems that if there is any impact from the use of the structured abstracts, it has been to reduce the likelihood of low scoring abstracts. Looking at all four plots in Figure 7, it appears that the spread of values as well as the central locations differ. Overall, the distribution of the data in each of the four plots do not look normally distributed and the change in distribution implies that the variances may not be the same in the different groups.

These issues (change in distribution, non-normal data and possible non-stable variance) suggest that we need to consider a robust analysis. There are three possible methods of analysing the data:

- Trimmed mean analysis of variance testing a linear combination of the trimmed means.
- ANOVA-like rank-based analysis testing the interaction term.
- Cliff's method adapted for differences in differences.

Using a trimmed mean factorial analysis, the mean values for completeness are shown in Table 8 and the results are:

- The Time period effect is significant (p=0.006)
- The Journal effect is not significant (p=0.062)
- The Interaction effect is not significant (p=0.065)

Table 8: Trimmed Means for Phase 1 Abstract Completeness

	IST	JSS
Period 1	0.5104	0.5097
Period 2	0.6711	0.5439

Testing the linear contrast directly gives an effect size of 0.1265 with 95% confidence limits (-0.008293 to 0.2613). The confidence interval spans zero so the effect size is not statistically significant at the $p = 0.05$ level.

In the past, it has not been possible to do non-parametric rank based tests for such complex designs. However, the more recent approach to rank-based ANOVA (which are also robust to problems associated with tied values and variance heterogeneity in ranks) does allow such an analysis (Akritas et al, 1997). Akritas et al.'s paper is extremely complicated, but fortunately, the procedure has been automated and is available in an **R** procedure, see (Wilcox, 2012, p. 260-261). Applying this analysis to our data gives the following results:

- The Time period effect is statistically significant with $p = 0.00091$.
- The Journal effect is statistically significant with $p = 0.0153$.
- The interaction is not statistically significant with $p = 0.10939$.

These results indicate that the completeness of the abstracts is better for the more recent studies and that the interaction term is not significant, which agree with the trimmed mean analysis. However, in contrast to the trimmed mean analysis, the rank-based study suggests that there is a significant journal effect.

As discussed in Section 3.3, when applying the ANOVA-like rank method to a design that is more complex than a simple two group experiment, the relative effect size is calculated as shown in Table 9.

However, the relative effect sizes do not consider the differences in differences effect (that is, they are exactly the same values that would be obtained if

Table 9: Relative Effect Sizes for Phase 1 Abstract Completeness

	IST	JSS
Period 1	0.4238	0.3731
Period 2	0.7219	0.4812

the data were treated simply as coming from a one factor experiment with four levels), so cannot act as an effect size for meta-analysis purposes. Without an effect size and the effect size variance, we cannot incorporate data from other independent studies using meta-analysis. For that reason we consider another analysis approach, based on Cliff's δ.

Using Cliff's δ, it is straightforward to test the impact of Time period and Journal on abstract completeness by performing two separate tests and ignoring the interaction term. The results of this analysis show:

- The Time period effect shows Time period 2 completeness exceeds Time period 1 completeness with $\delta = 0.4065$ and 95% confidence interval (0.1581 to 0.6061)
- The Journal effect shows that IST completeness exceeds JSS completeness with $\delta = 0.2912$ and 95% confidence interval (0.02908 to 0.5159).

Cliff's approach is not restricted to two samples. He pointed out that confidence interval of the difference between two independent δ's can also be assessed (see Cliff, Equation 19) as follows:

$$(d_2 - d_1) \pm z_{\frac{\alpha}{2}} \left(s_{d_2}^2 + s_{d_2}^2\right)^{\frac{1}{2}} \tag{34}$$

This is exactly what we need for a difference in differences analysis. In this case d_2 is the difference between the probability that IST abstracts score higher than JSS abstracts in period 2 and the probability that JSS abstracts score higher than IST abstracts in period 2 and d_1 is the equivalent difference for period 1. In effect, we reduce the d-value obtained for the period 2 observations to account for difference between the groups in period 1 (the control situation). Using Wilcox's algorithms to calculate the d values for each time period and calculating the differences in differences statistics ourselves, our data gives the result shown in Table 10.

Table 10: Cliff's d for Phase 1 Abstract Completeness

	Period 1	Period 2	Difference
p_1	0.5075	0.735	
p_2	0.0425	0.02	
p_3	0.45	0.245	
d	0.0575	0.49	0.4325
s_d	0.0374	0.0278	0.0465

Since $z_{\frac{\alpha}{2}} = 1.96$, the 95% confidence interval for the difference of the differences is (0.3413, 0.5237). Because the effect size is positive and the confidence interval does not include zero, the difference in difference analysis based on Cliff's δ suggests that the IST abstracts are more complete than JSS abstracts after the introduction of structured abstracts, after allowing for the fact that the IST abstracts were slightly more complete than the JSS abstracts before the introduction of structured abstracts. This result is inconsistent with the results found by the trimmed mean analysis and the ANOVA-like rank-based method. However, for the purposes of this example we will continue to use Cliff's approach.

We obtain similar results when viewing the kernel density plots of the clarity scores for each group (see Figure 8) and analysing the clarity data (see Table 11). In this case, the effect size is estimated as 0.25 with 95% confidence interval $(0.1159, 0.3091)$. A similar result is to be expected because, as shown in Figure 9, abstract completeness and clarity are highly correlated (Kendall's $tau = 0.518, p < 0.0001$).

Table 11: Ciff's d for Phase 1 Abstract Clarity

	Period 1	Period 2	Difference
p_1	0.46	0.5875	
p_2	0.1175	0.075	
p_3	0.4225	0.3375	
d	0.0375	0.25	0.2125
s_d	0.036	0.0336	0.0493

Phase 2 Analysis

In this section we analyse the data from Wroclaw University of Science and Technology and discuss how it can be aggregated with the previous data. As previously noted, the second set of data from Hong Kong showed no evidence of consensus about abstract complexity and clarity, so could not be used.

Analysing the data in isolation, the results for the Brunner et al. relative effect sizes for completeness were similar to those found in the first period (see Table 12), but the effects were not significant:

- The Time period effect is not statistically significant with $p = 0.41535$.
- The Journal effect is not statistically significant with $p = 0.6621164$.
- The interaction is not statistically significant with $p = 0.44237$.

Cliff's difference of differences analysis is shown in Table 13 (completeness) and Table 14 (clarity).

For completeness, the standard error is large enough to indicate that the effect size is not statistically significant. Furthermore, for clarity the effect size is negative. Thus, analysed by itself the Polish data does not support the

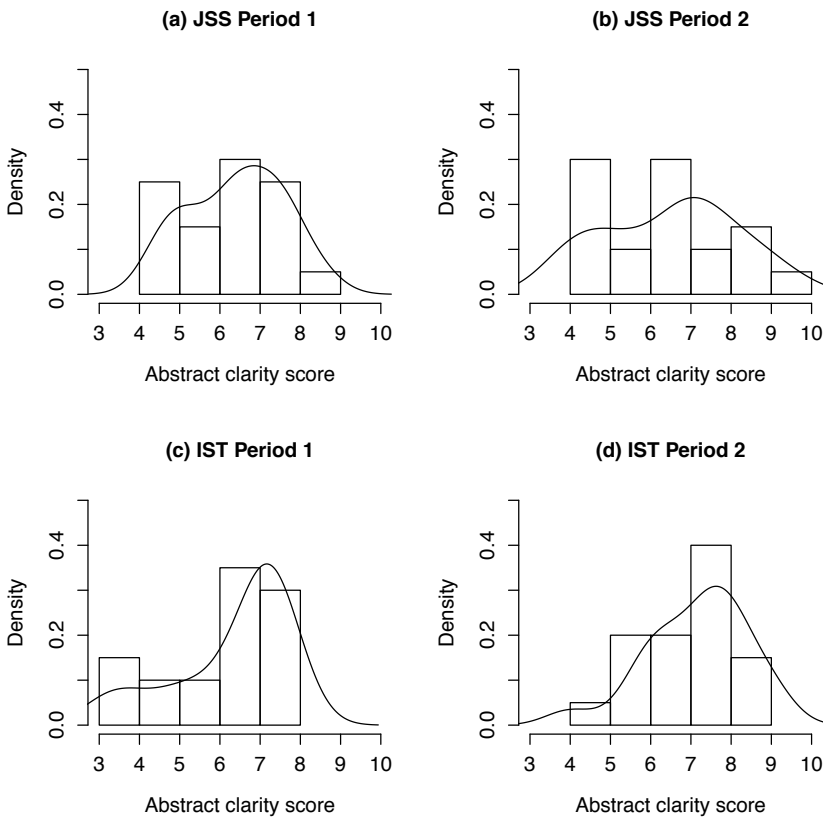

Fig. 8: Kernel density plots of the median clarity score for each abstract

Table 12: Relative Effect Sizes for Phase 2 Abstract Completeness

	IST	JSS
Period 1	0.4062	0.4609
Period 2	0.6641	0.4688

hypothesis that structured abstracts improve completeness and clarity. The number of abstracts is clearly insufficient to provide statistically significant results and estimates of d have a large standard error.

The correct way to incorporate the results of data collected after the analysis of an initial tranche of data is via meta-analysis (Braver et al, 2014). Just adding the new data to the existing data set is wrong, since it involves deciding to collect more data after looking at the results (John et al, 2012). Equally, meta-analysis of all six studies is not a valid approach because the five

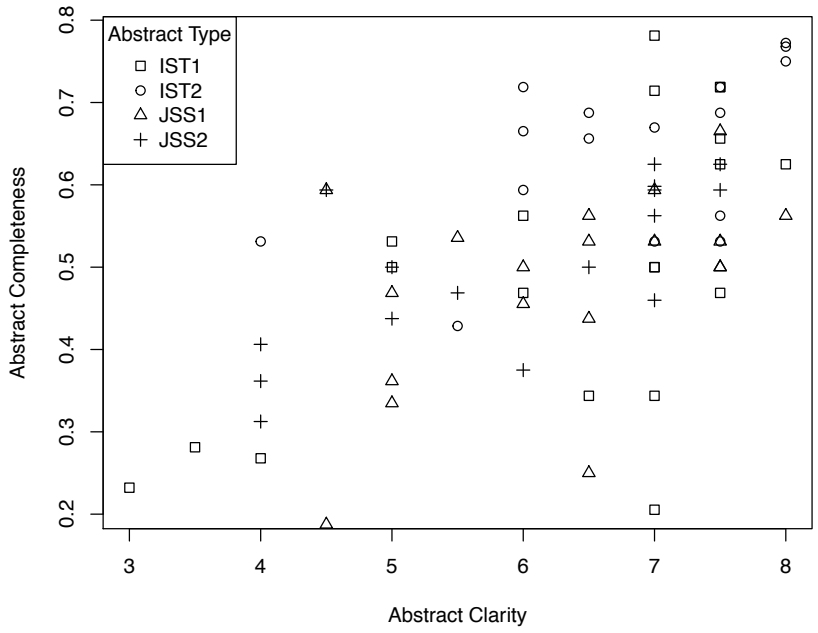

Fig. 9: Scatter plot of Abstract Clarity vs Completeness

Table 13: Cliff's d for Phase 2 Abstract Completeness

	Period 1	Period 2	Difference
p_1	0.5	0.6875	
p_2	0	0	
p_3	0.5	0.3125	
d	0	0.375	0.375

studies in the first tranche were planned in advance (before the experiment) as defined in the protocol. Thus, they are treated as one distributed experiment.

When undertaking a meta-analysis, it is important to decide whether to perform a fixed-effects analysis or a random-effects analysis. Borenstein et al (2009) discuss whether meta-analysts should use a fixed-effect or a random-effect analysis. They suggest a fixed effects analysis is appropriate if two conditions are met. Firstly the analysts believe that all the studies are functionally similar, secondly the goal is to compute the common effect size for the identical population, and not to generalise to other populations. In our

Table 14: Ciff's d for Phase 2 Abstract Clarity

	Period 1	Period 2	Difference
p_1	0.8125	0.6875	
p_2	0.125	0.0625	
p_3	0.0625	0.25	
d	0.75	0.4375	−0.3125

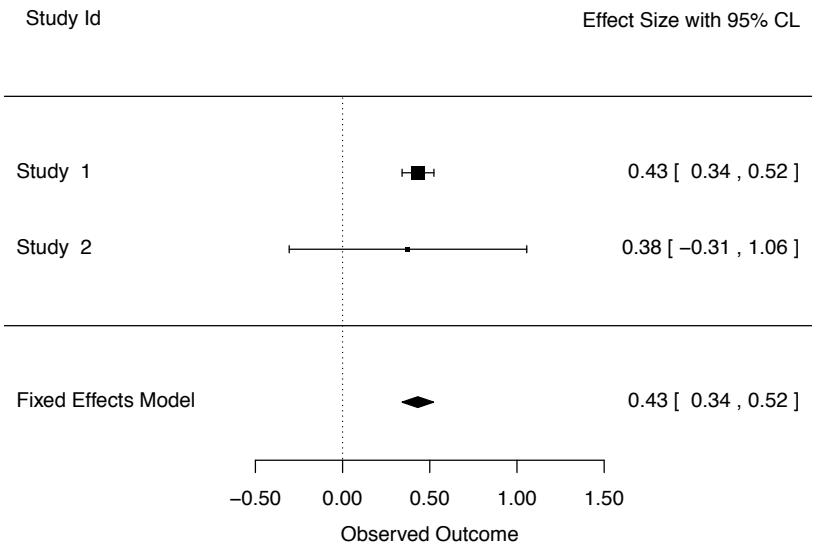

Fig. 10: Forest plot of Completeness results from two studies

case the use of exactly the same protocol and output variables model and the limited goal of our meta-analysis suggest that a fixed-effect size is justified.

Using the **R metafor** package (Viechtbauer, 2010) and a fixed effects analysis, the aggregated effect size, for the completeness data, was estimated to be 0.4315 with 95% confidence interval $(0.3411, 0.5219)$.

The forest plot of the analysis is shown in Figure 10. The meta-analysis indicates that the second data set has no effect on the results obtained from analyzing the first data set. Braver et al. make it clear that this is what is likely to happen with a low power replication.

For the clarity data, the effect size is reversed, and the fixed effects analysis showed evidence of heterogeneity ($Q = 4.6908, df = 1, p = 0.0303$). This suggests that the clarity data results from each data collection period should not be aggregated into an overall effect size using a fixed-effects model.

4.4 Discussion of the multi-site example

An important issue arising from the multi-site experiment is that analyses performed using the different non-parametric methods gave different results. We must reemphasize that we do *not* advocate trying every possible method of analysis until finding one that gives a significant result. There needs to be a good reason for rejecting or selecting a specific analysis method.

Since the completeness and clarity metrics used in our experiment are both restricted—completeness to between 0 and 1 and clarity to between 0 and 10, and because the kernel density plots look as if the impact of structured abstracts is to reduce the likelihood of very incomplete abstracts, we would expect the non-parametric analyses to be more reliable than the trimmed mean analysis.

The contrast between the analysis using Cliff's δ and the rank-based ANOVA-like analysis may have occurred because the rank method:

- Uses ranks obtained across all groups, which may reduce the rank differences between specific groups.
- Includes the mid-rank values used to cater for tied values while Cliff's method removes the impact of tied values.
- Uses statistical tests that allow for variance heterogeneity between groups but that result in a reduction in the degrees of freedom for the F test.

Also, within the context of simple two group comparisons, Wilcox (2012) suggests that, with many tied values, Cliff's method may be a bit better than the rank-based method in terms of achieving a Type 1 probability less than the nominal alpha level. However, he does not discuss the impact of more complex statistical designs.

It may also be because the effect size is relatively small. Although a δ value of 0.4325 would be considered large according to Table 5, the effect is rather small compared with that obtained in a previous controlled experiment. The median of the 20 median abstract completeness scores for each experimental condition is shown in Table 15.

Table 15: Median Abstract Completeness

	Period 1	Period 2
IST	0.5	0.6786
JSS	0.5156	0.5781

This suggests that the median score is increased by approximately 0.1 which is equivalent to getting one additional *Yes* answer in the 8 completeness questions. In contrast, Budgen et al (2011) observed a median difference between conventional and structured abstracts of just over 0.2 using a similar scoring method. In addition, the abstracts in the Budgen study were all written by undergraduates who would have had little experience of writing abstracts, whereas the abstracts in our multi-site experiment were written by the authors of the papers. Authors, even if they were post graduate students, would be more experienced than computer science undergraduates. Thus, the likely impact of using structured abstracts would be greater in the previous study.

However, all of these are post-hoc justifications and we would be cautious about claiming that we can reject the null hypothesis that there is no difference between structured and conventional abstracts. It would be useful to replicate the experiment with data from a different set of abstracts. We note that such a replication should:

- Keep the title and keywords with the abstract on the evaluation form to be more consistent with research practice.
- Ensure that abstracts selected from JSS and IST for study should come from the same two time periods. The observed increase on completeness between the time periods suggests that we should have ensured that the time periods for both JSS and IST were exactly the same. As it is, there is a risk that any conventional IST abstracts obtained from the last six months of 2010 would have a greater completeness value than earlier conventional abstracts. This would have lead to an increased average completeness for IST period 1 abstracts, which would have reduced the likelihood of detecting a difference in differences effect.
- Review the evaluation questions themselves to see whether they can be made more objective. The lower levels of agreement among judges who do not have English as a first language may be a result of the abstracts, but could also be due to problems with the evaluation questions.

Our experimental design is not appropriate for testing hypotheses regarding overall time-trends in abstract completeness. However, our results suggest that the overall quality of abstracts has improved in the second time period for both JSS and IST. This could be explained because general criticisms of abstracts in systematic literature reviews, together with experimental results suggesting structured abstracts were likely to be more complete than conventional abstracts, would probably have increased awareness of the need for good quality abstracts and helped produce an overall improvement. However, to properly test the hypothesis of a general improvement an experiment would need to test the completeness of abstracts across a wide range of journals.

In terms of advantages of non-parametric methods, looking at Figure 7, the multi-site example, suggests that the new non-parametric methods are preferable to conventional analysis methods because they are able to detect changes related to the overall distribution, not just the mean.

5 Discussion

This section summarises arguments in favour of the use of robust statistics and identifies limitations associated with their use.

5.1 Arguments for the use of robust statistics

We have proposed using analysis techniques that are robust to non-normality when we have reason to believe our data is non-normal. We have also suggested the use of Kernel density functions to identify empirical distributions that appear non-normal. However, we have not discussed whether we should use quantile-quantile plots (q-q plots) or statistical tests to check for normality, nor have we discussed whether it is preferable to transform data.

With respect to q-q plots, like kernel density plots, they require the analyst to make a judgment about whether the data is normal (or normal enough) or not. In our view, the kernel density plots are somewhat easier to interpret, but we accept that this is a matter of personal preference.

With respect to tests for normality, we note that advocates of robust statistical methods usually state that tests of normality have poor power (see (Wilcox, 2012) or (Erceg-Hurn and Mirosevich, 2008)), whereas advocates of the normality tests publish papers demonstrating that their tests achieve good power (see for example, (D'Agostino et al, 1990), (Mudholkar et al, 2002), or (Shapiro et al, 1968)). In a more recent study, Razali and Wah (2011) compared the Shapiro-Wilk test, the Kolmogorov-Smirnov test, the Lillefors test and the Anderson-Darling test and concluded that the Shapiro-Wilk test is the most powerful normality test, but that the power of all four tests was low for small sample sizes (that is, sample sizes of 30 and below). The Shapiro-Wilk test results for the data sets discussed in this paper are shown in Table 16.

The table shows that the Shapiro-Wilk test suggests more of the data sets are normally distributed than inspection of the Kernel density plots would indicate. In addition, if we use normality tests and they suggest some groups have normally distributed data and some do not, applying a transformation to all groups (which is necessary for any valid statistical analysis) may reduce the normality of any group which had more or less normal data to begin with. Overall, with relatively small, messy data sets it seems best to err on the side of caution and assume that the data is non-normal. Under such circumstances adopting robust methods may sometimes be conservative, but using non-robust analysis methods would make the results of any analysis untrustworthy.

With respect to using transformations to make data sets more Normal, our example suggests that the logarithmic transformation is not a panacea for all data sets. It is also the case that the use of standard transformations (such as logarithms or square roots) can make interpreting results more difficult. In addition, Wilcox and Keselman (2003) point out that:

- Simple transformations do not guard against low statistical power when dealing with heavy-tailed distributions.

Table 16: Shapiro-Wilk Normality Test probability for example data sets (data sets available from the **reproducer R** package (Madeyski, 2015))

Data Set	Measure	Data Set Size	p-value of test	p-value for log transformed data
Finnish Data	Effort	38	0.0004	0.0653
Software Defect Prediction				
Simple Model	% Modules	34	0.3917	0.0123
Advanced Model (NDC)	% Modules	34	0.0373	0.0101
COCOMO				
Embedded	Productivity	28	<0.0001	0.7734
Semi-Detached	Productivity	12	0.6135	0.0161
Organic	Productivity	23	0.0379	0.7103
Abstract Experiment Data				
JSS1	Completeness	20	0.0899	0.0017
JSS2	Completeness	20	0.4853	0.8029
IST1	Completeness	20	0.3194	0.0658
IST2	Completeness	20	0.6371	0.2059

– Simple transformations can alter skewed distributions but do not deal directly with outliers.

They recommend the use of trimmed means as an effective transformation method for heavy tailed distributions. In this paper, we have recommended the use of trimmed means when we are concerned about changes to the central location of a data set. In addition, trimmed means are justified in a number of different ways:

– They are a compromise between the median (maximum trimming) and the mean (zero trimming).
– They are a form of weighted mean.
– They are based on excluding the observations that provide least information about the central location.
– They are in common use for scoring competitions where performance and style are judged subjectively, for example, scoring diving competitions where the two upper and lower values from seven assessments are discarded.

The other general analysis approach that can be used with non-Normal data is robust non-parametric analysis. We have discussed the need for newer non-parametric tests, in particular the ANOVA-like rank-based method developed by Akristas, Brunner and colleagues, and Cliff's δ. The advantages of the new forms of non-parametric metrics and of tests based on those metrics are:

– They are the best way of testing ordinal scale measures. In Software Engineering many of our measures (other than those related to elapsed time) have no physical basis, and are more likely to be ordinal than interval or ratio measures. For example, function points and any measures

constructed primarily from subjective assessments. This includes metrics such as the abstract completeness score used in our example in Section 4.
- $\hat{P}$ and δ provide sensible non-parametric effect sizes. Indeed for meta-analysis, Kromrey et al (2005) report that Cliff's δ outperformed Cohen's d and Hedges g statistics.
- For purposes of meta-analysis studies, it is possible to convert the MWW U or the Wilcoxon W statistic into $\hat{P}$ or δ. Although it should be noted that there is some disagreement about terminology. For example, **R** reports the W statistic (that is, the sum of the ranks of the first group) but labels it U.
- $\hat{P}$ and δ do not suffer from the large scale approximation problems associated with U or W.
- Brunner's and Cliff's methods are implemented in **R** source code provided by Wilcox.
- Both methods can be used with more complex designs than simple between-groups designs, including repeated measures designs. The rank-based ANOVA-like approach can be applied to virtually all standard experimental designs, including n by m factorials.

Furthermore, both approaches have been adopted in recent published software engineering studies. Cliff's δ was used in (El-Attar, 2014), (El-Attar et al, 2012) and (Tappenden and Miller, 2014). The probability of superiority metric $\hat{P}$ (referred to as $\hat{A_{12}}$) was used in (Madeyski et al, 2014, 2012).

5.2 Limitations of robust statistical methods

If data sets are normally distributed or sample sizes relatively large, the robust methods are less powerful than the standard methods. However, in many cases, the robust methods are designed to be reasonably powerful even if the data are normal, and they are considerably more powerful if the data are not normally distributed or sample sizes are small.

A related issue is that all the robust methods discussed in this paper (with the exception of Cliff's method) will lead to a reduction in the degrees of freedom available for statistical tests and the construction of confidence intervals. For the parametric methods, trimming which removes large and small data values, and the use of Welch's test both contribute to a reduction in the degrees of freedom (compare, for example, the number of projects in each Mode type in Table 3 with the degrees of freedom for the trimmed mean statistical test shown in Table 4).

Finally, the use of power analysis to estimate required sample size is more complex for robust methods. In particular, the relationship between degrees of freedom and the group variances in Welch's test (see equation 16) complicates any power analysis for trimmed means or the rank-based ANOVA-like method.

6 Conclusions

Classical statistical analysis methods have limitations when dealing with real data that are skewed, and/or heavy-tailed, and/or have unstable variances. Box plots can also conceal the extent of non-normality. We recommend using kernel density plots to inspect the distribution of data.

Parametric tests such as t and F tests are not robust to non-normality, particularly severe skewness and combinations of non-normal properties. For comparing the central location of different data sets, we recommend using Yuen's test based on trimmed means and Welch's test for unequal variances.

Rank-based methods such as MWW and Kruskal-Wallis have problems *when statistical tests are based on large sample approximations* for the rank variance. Furthermore, since the U and W test statistics are based on rank averages which increase as the number of observations increase, they do not deliver reliable effect sizes. For analyses that are concerned with general shifts in the distribution rather than changes in the central location or are concerned that their data are naturally ordinal-scaled, we recommend using Cliff's or Brunner et al.'s methods for robust non-parametric methods with Cliff's δ or the probability of superiority as effect sizes.

Appendices

A The theory of M-estimators

This section gives a very brief introduction to the theory of M-estimators based primarily on (Goodall, 1983).

The mean of a sample can be considered as the value that minimises the sum of squared deviations from itself. Equivalently, the variance can be considered as the function of a sample which penalises observations that differ from the mean. These ideas are shown in the top two graphics in Figure 11. Figure 11 (a) shows the variance function assuming a symmetric distribution about a central location t. In the context of M-estimators the function that describes the deviations from the central value is called the *objective function*. The objective function makes it clear that the expected value of the squared deviation of an observation x from t (that is, $E(x-t)^2$) gives greatest weight to values that are far from t. In other words, the function $(x-t)^2$ increases rapidly as x move away from t.

The derivative of the objective function is called the $influence function$. The influence function for the mean is shown in Figure 11 (b). The influence function passes through 0 at the point t confirming that t is the value that minimises the objective function. Furthermore, it shows that the central location is unbounded and if one value x increases towards infinity, the mean value of the sample will likewise increase towards infinity. This is a slightly more formal way of explaining why the mean is not a robust measure of central location.

Figure 11 (c) and (d) show the objective function and influence function of the median respectively. The objective function of the median is defined to be:

$$\rho\left(x,t\right)=\left|x-t\right| \tag{35}$$

with the corresponding influence function:

$$\psi\left(x,t\right)=sign(u) \tag{36}$$

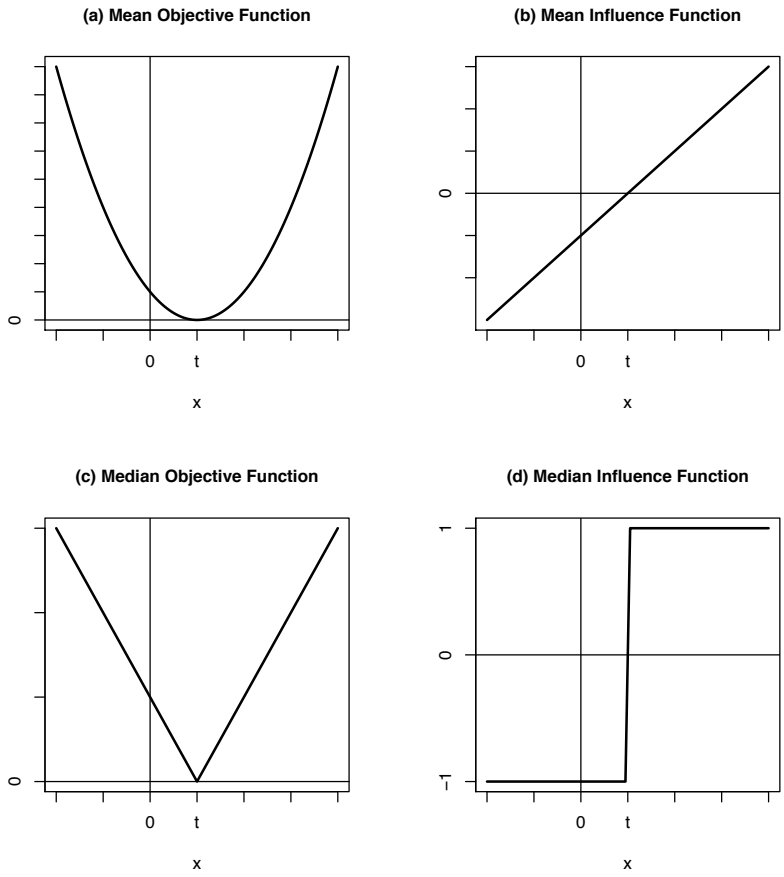

Fig. 11: The Objective Functions and Influence Functions of the Mean and Median

$$sign(u) = \begin{cases} +1 \text{ if } u > 0 \\ 0 \quad \text{if } u = 0 \\ -1 \text{ if } u < 0 \end{cases} \tag{37}$$

The objective function of the median in Figure 11 (c) can be seen to penalise values close to the central location more heavily than the squared deviations, but does not penalise distant points as severely. Furthermore, the influence function of the median shown in Figure 11 (d) confirms that the median is bounded and thus unaffected by a single value increasing towards infinity.

If we regard the objective function as simply a method of weighting deviations from the central location there is no reason to restrict ourselves to the squared deviation or the absolute deviation. We can choose any objective function or influence function that has desirable properties. For example, Huber's objective function corresponds to an absolute

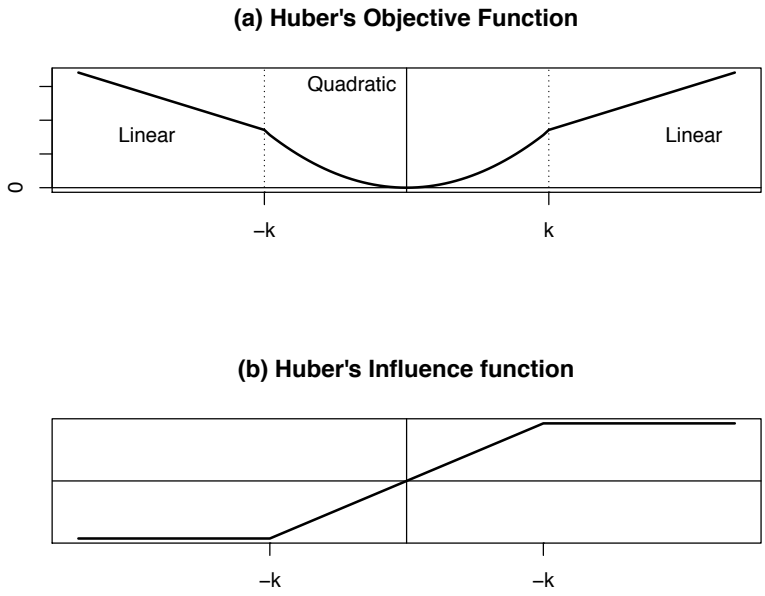

Fig. 12: Huber's Objective and Influence Functions

deviation at the extremes of the distribution and the squared deviation near the centre of the distribution. Huber's objective function and influence function are shown in Figure 12.

Figure 12 (b)) shown that the influence function is linear at the centre so the central value does not change abruptly like the median, but is constant in its tails, so the central value is not unstable if there are a few unusually large or small values.

Huber's objective function is:

$$\rho = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < k \\ k|x| - \frac{1}{2}x^2 & \text{if } |x| \geq k. \end{cases} \tag{38}$$

with influence function:

$$\psi = \begin{cases} x & \text{if } |x| < k \\ ksign(x) & \text{if } |x| \geq k. \end{cases} \tag{39}$$

Huber's objective function can be used to estimate a central location value which is the most efficient for a contaminated Gaussian distribution (that is, a distribution where the majority of the distribution comes from one Gaussian distribution but a small percentage comes from another Gaussian distribution with a much larger mean or standard deviation). The above discussion provides the context for formally defining an M-estimate.

Definition: The M-estimate $T_n(x_1, ..., x_n)$ for the function ρ is the value of t that minimizes the objective function $\Sigma_{i=1}^{n}\rho(x_j, t)$.

The properties of ρ and ψ determine the properties of a specific M-estimator. Goodall discusses the properties that are appropriate for a *robust* M-estimator of the central location:

1. The *breakdown bound* of the estimator should be large. The breakdown bound is the largest possible fraction of the observations for which there is a bound on the estimate

when that fraction is altered without restriction. For example, in the case of the mean, as a single observation is made larger, the mean increases without bound. Therefore, the breakdown bound of the mean is 0. For the median, the breakdown bound is $1/2 - 1/n$ for even n and $1/2 - 1/(2n)$ for odd n. For Huber's central location estimator, the breakdown bound is $p(k)$ where $p(k)$ is the proportion of the data set greater than k. For a trimmed mean, the breakdown bound corresponds to the extent of trimming, where a trimming constant of $\gamma = 0.2$ causes 40% of the sample values to be trimmed.

2. The estimate should have finite *gross error sensitivity* which means that the influence function is bounded. Gross error sensitivity is the maximum effect a contaminated value can have on the M-estimator, which corresponds to then maximum absolute value of the influence function. For the median, Huber's central location estimator, and the trimmed mean, the gross error sensitivity is bounded but not for the mean.
3. The estimate should have *finite local-shift sensitivity*. This means that the influence function should not have large discontinuities as can be the case for the median. The Winsorized and trimmed means also exhibit discontinuities related to the point in the data where the x-values are replaced either by the upper and lower γ-percentile or by 0 respectively.
4. The estimator should be resistant to very large outliers. This is one of the most important features of a robust estimator. In terms of the influence function, it means it must have a *finite rejection point*. The rejection point is the least distance from the location estimator beyond which observations do not contribute to the value of the estimate. The mean, median and Huber's central location estimator do not have finite rejection points, but the trimmed mean and Winsorised mean do.
5. Some M-estimators are *maximum likelihood estimators* which by definition maximise the likelihood of getting the observed data. There is a connection between the influence function and the underlying density of a distribution. So, if we are sure of the underlying distribution, we can select an influence function that will allow us to calculate an *asymptotically efficient estimator* for that distribution.
6. For Gaussian data the influence function of the mean is linear. Since given the central limit theory, all distributions are approximately normal at their centre, we might want an influence function that is linear near the centre, which means that:

$$\psi(x) \approx kx \tag{40}$$

where k is non-zero constant usually standardised to equal 1.

7. For a symmetric underlying distribution, the objective function of an M-estimator should give equal weight to observations at equal distances either side of the centre. This means that the influence function should be *odd*, so that:

$$\psi(-x) = \psi(x) \tag{41}$$

Researchers have derived robust measures of central location and scale based on developing influence functions that have the desirable properties itemized above. Figure 13 shows two such functions, the biweight influence function (also known as Tukey's bisquare), which has the equation:

$$\psi = \begin{cases} x(1-x^2)^2 & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1 \end{cases} \tag{42}$$

and the Andrews influence function, which has the equation:

$$\psi = \begin{cases} sin(x/a) & \text{if } |x| \leq a\pi \\ 0 & \text{if } |x| > a\pi. \end{cases} \tag{43}$$

Both influence functions are fairly linear about the origin and tend to zero at the extreme values. Wilcox (2012) discusses both the use of the biweight and Huber's ψ in the construction robust measures of scale, and how such measures are used to support robust regression. However, these issues are beyond the scope of this paper.

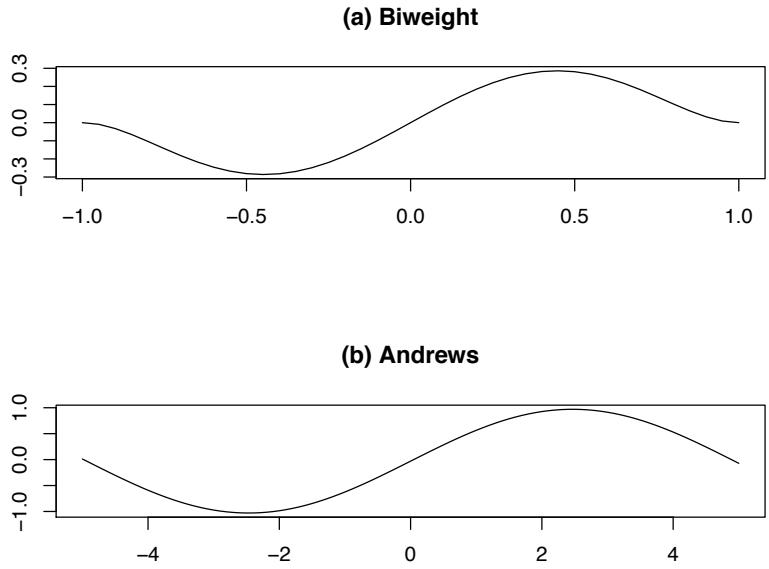

Fig. 13: The Biweight and Andrew's Influence Functions

B Overview of the Multi-Site experiment

Figure 14 shows an E-R style diagram giving an overview of the experiment. An explanation of the entities in the model follows.

Abstract An **Abstract** was selected at random from one **Source**. **Abstracts** are selected without replacement, so an **Abstract** was assigned to a specific **Site** and no other site (with the exception that sites 6 and 7 used the same set of abstracts). Each **Abstract** was assessed by 4 **Judges** and so has 4 associated **Evaluation Forms**.

Source The origin of the **Abstract**, being one of two journals (IST or JSS), in one of two specific time periods.

Site A specific educational institution taking part in the experiment. In the first data collection phase the sites were: Keele University (UK), Durham University (UK), PSU (Thailand), Lincoln University (NZ), Polytechnic University (HK). In the second data collection phase the sites were: City University (HK), Wroclaw University of Science and Technology (Poland). **Sites** were each assigned 16 **Abstracts**.

Judge Each **Site** recruited 16 **Judges** intended to be students at the end of their second year of CS/IT studies. Each **Judge** evaluated 4 of the 16 **Abstracts** assigned to his/her **Site**.

Assessment Group An **Assessment Group** comprised 4 **Judges** who all assessed the same 4 **Abstracts**.

Sequence Order **Judges** in a specific **Assessment Group** assessed **Abstracts** in one of two balanced **Sequence Orders**. Each **Judge** in a specific **Assessment Group** was intended to assess the 4 **Abstracts** assigned to the **Assessment Group** in a different sequence (see Table 17, where JSS-1 means an abstract from JSS time period 1, JSS-2 an abstract from JSS in time period 2, IST-1 an abstract from IST in time period 1

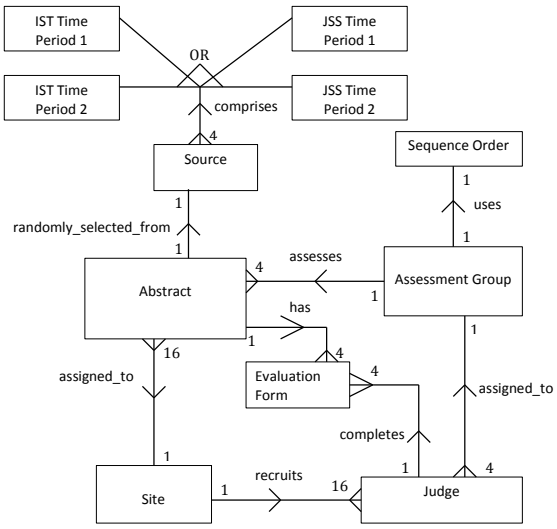

Fig. 14: Diagram illustrating the organisation of the experiment

and IST-2 an abstract from IST in time period 2). Each **Site** was required to use the given sequence for each pairing of four abstracts and four judges. Although the protocol required adopting the different sequences, in practice, some sites did not conform with the order process (see (Budgen et al, 2013)).

Evaluation Form Each **Evaluation Form** contained the completeness and clarity information provided by a specific **Judge**.

Table 17: Sequence Order for Viewing Abstracts for Four Judges and Four Abstracts

Judge	First	Second	Third	Fourth
J1	JSS-1	JSS-2	IST-2	IST-1
J2	JSS-2	IST-1	JSS-1	IST-2
J3	IST-1	IST-2	JSS-2	JSS-1
J4	IST-2	JSS-1	IST-1	JSS-2

Figure 15 defines the process used to select the abstracts for the experiment and construct the experimental materials.

Figure 16 is a flow diagram showing a high-level overview of the experimental process undertaken at each site.

Figure 17 is flow diagram of the experimental process from the viewpoint of the judges.

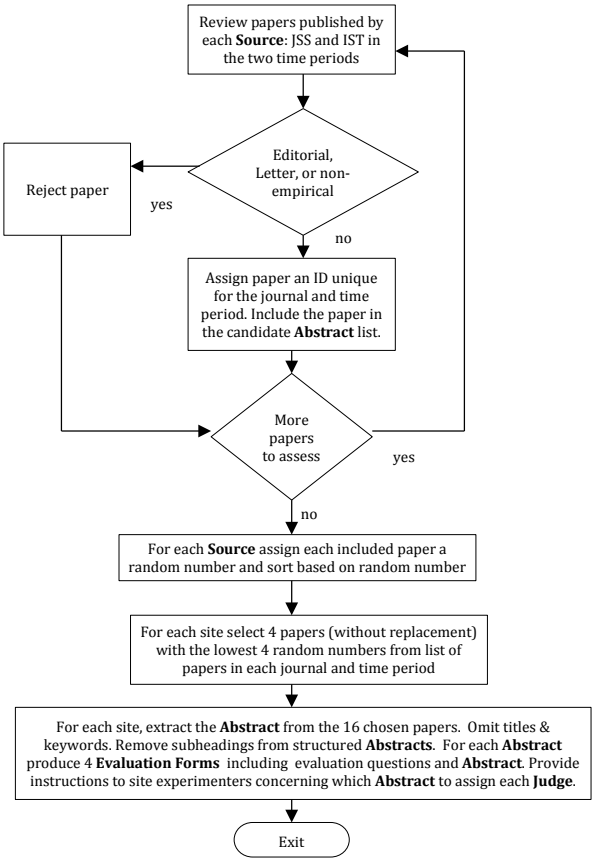

Fig. 15: Flow diagram of the Experimental Material Preparation Process

References

Acion L, Peterson JJ, Temple S, Arndt S (2006) Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. Statistics in Medicine 25(4):591–602, DOI 10.1002/sim.2256

Agresti A, Pendergast J (1986) Comparing mean ranks for repeated measures data. Communications in Statistics - Theory and Methods 15(5):1417–1433

Akritas MG, Arnold SF (1994) Fully Nonparametric Hypotheses for Factorial Designs I: Multivariate Repeated Measures Designs. Journal of the American Statistical Association 89(425):336–343, DOI 10.1080/01621459.1994.10476475

Akritas MG, Arnold SF, Brunner E (1997) Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs. Journal of the American Statistical Association 92(437):258–265, DOI 10.1080/01621459.1997.10473623

Arcuri A, Briand L (2011) A practical guide for using statistical tests to assess randomized algorithms in software engineering. In: ACM/IEEE International Conference on Software Engineering (ICSE), IEEE, pp 1–10, DOI 10.1145/1985793.1985795

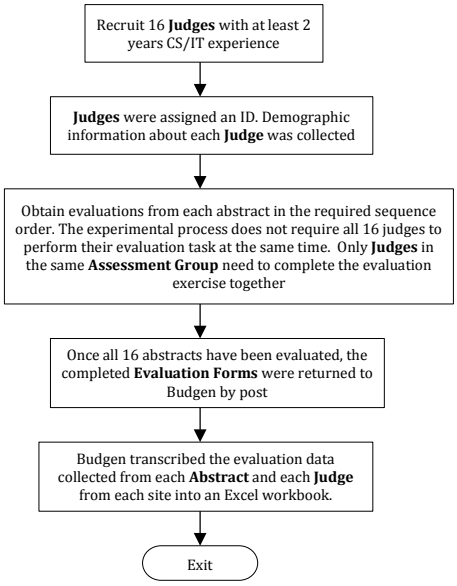

Fig. 16: Flow diagram of the Experimental Process at each Site

Arcuri A, Briand L (2014) A hitchhiker's guide to statistical tests for assessing randomized algorithms in software engineering. Software Testing, Verification and Reliability 24(3):219–250, DOI 10.1002/stvr.1486

Behrens JT (1997) Principles and procedures of exploratory data analysis. Psychological Methods 2(2):131–160

Bergmann R, Ludbrook J, Spooren WPJM (2000) Different Outcomes of the Wilcoxon-Mann-Whitney Test from Different Statistics Packages. The American Statistician 54(1):72–77

Boehm BW (1981) Software Engineering Economics. Prentice-Hall

Borenstein M, Hedges LV, Higgins JP, Hannah R R (2009) Introduction to Meta-Analysis. John Wiley & Sons Ltd.

Box GEP (1954) Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification. The Annals of Mathematical Statistics 25(2):290–302, DOI 10.1214/aoms/1177728786

Braver SL, Thoemmes FJ, Rosenthal R (2014) Continuously cumulating meta-analysis and replicability. Perspectives on Psychological Science 9(3):333–342, DOI 10.1177/1745691614529796

Brunner E, Munzel U, Puri ML (2002) The multivariate nonparametric Behrens–Fisher problem. Journal of Statistical Planning and Inference 108(1–2):37–53, DOI 10.1016/S0378-3758(02)00269-0

Budgen D, Kitchenham BA, Charters SM, Turner M, Brereton P, Linkman SG (2008) Presenting software engineering results using structured abstracts: a randomised experiment. Empirical Software Engineering 13(4):435–468, DOI 10.1007/s10664-008-9075-7

Budgen D, Burn AJ, Kitchenham B (2011) Reporting computing projects through structured abstracts: a quasi-experiment. Empirical Software Engineering 16(2):244–277, DOI 10.1007/s10664-010-9139-3

Budgen D, Kitchenham B, Charters S, Gibbs S, Pohthong A, Keung J, Brereton P (2013) Lessons from conducting a distributed quasi-experiment. In: 2013 ACM / IEEE

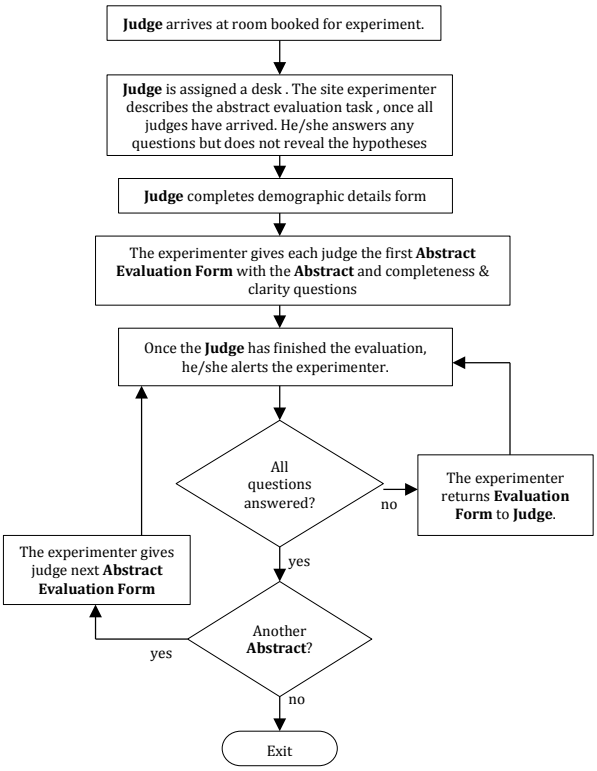

Fig. 17: Flow diagram of the Experimental Process from the Judge's viewpoint

International Symposium on Empirical Software Engineering and Measurement, pp 143–152, DOI 10.1109/ESEM.2013.12

Cliff N (1993) Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin 114(3):494–509

Cohen JW (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, New York, USA

Cohen JW (1992) A power primer. Psychological Bulletin 112(1):155–159

Conover W, Imam RL (1981) Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician 35(3):124–129

D'Agostino RB, Belanger A, D'Agostino J Ralph B (1990) A suggestion for using powerful and informative tests of normality. The American Statistician 44(4):316–321

Dejaeger K, Verbeke W, Martens D, Baesens B (2012) Data mining techniques for software effort estimation: A comparative study. IEEE Transactions on Software Engineering 38(2):357–397

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:1–30

Dybå T, Kampenes VB, Sjøberg DIK (2006) A systematic review of statistical power in software engineering experiments. Information and Software Technology 48(8):745–755, DOI 10.1016/j.infsof.2005.08.009

El-Attar M (2014) Using SMCD to reduce inconsistencies in misuse case models: A subject-based empirical evaluation. Journal of Systems and Software 87:104–118, DOI 10.1016/j.jss.2013.10.017

El-Attar M, Elish M, Mahmood S, Miller J (2012) Is In-Depth Object-Oriented Knowledge Necessary to Develop Quality Robustness Diagrams? Journal of Software 7(11):2538–2552, DOI 10.4304/jsw.7.11.2538-2552

Erceg-Hurn DM, Mirosevich VM (2008) Modern robust statistical methods an easy way to maximize the accuracy and power of your research. American Psychologist 63(7):591–601

Gandrud C (2015) Reproducible Research with R and R Studio. CRC Press

Goodall C (1983) Understanding Robust and Exploratory Data Analysis. John Wiley and Sons Inc., chap M-Estimators of Location: An outline of the theory, pp 339–403

Grissom RJ (1996) The magical number .7 ± .2: Meta-meta-analysis of the probability of superior outcome in comparisons involving therapy, placebo, and control. Journal of Consulting and Clinical Psychology 64(5):973–982, DOI 10.1037/0022-006X.64.5.973

Hoaglin DC, Mosteller F, Tukey JW (eds) (1983) Understanding Robust and Exploratory Data Analysis. John Wiley & Sons, Inc.

Huijgens H, van Solingen R, van Deursen A (2013) How to build a good practice software project portfolio? Tech. Rep. TUD-SERG-2013-019, Delft University of Technology

John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 23(5):524–532, DOI 10.1177/0956797611430953

Jureczko M, Madeyski L (2015) Cross–project defect prediction with respect to code ownership model: An empirical study. e-Informatica Software Engineering Journal 9(1):21–35, DOI 10.5277/e-Inf150102

Kampenes VB, Dybå T, Hannay JE, Sjøberg DIK (2007) A systematic review of effect size in software engineering experiments. Information and Software Technology 49(11-12):1073–1086, DOI 10.1016/j.infsof.2007.02.015

Kitchenham B (1996) Software Metrics: Measurement for Software Process Improvement. Blackwell Publishers Inc.

Kitchenham B (2015) Robust Statistical Methods: Why, What and How: Keynote. In: Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (EASE 2015), pp 1:1–1:6, DOI 10.1145/2745802.2747956

Kitchenham B, Känsälä K (1983) Inter-item correlations among function points. In: Proceedings ICSE 15, IEEE Computer Society Press, pp 477–480

Kraemer HC, Kupfer DJ (2006) Size of Treatment Effects and Their Importance to Clinical Research and Practice. Biological Psychiatry 59(11):990–996, DOI 10.1016/j.biopsych.2005.09.014

Kromrey JD, Hogarty KY, Ferron JM, Hines CV, Hess MR (2005) Robustness in meta-analysis: An empirical comparison of point and interval estimates of standardized mean differences and Cliff's delta. In: Proceedings of the Joint Statistical Meetings, Minneapolis

Lipsey MW, Wilson DB (2001) Practical Meta-Analysis. Sage Publications, California, USA

Madeyski L (2010) Test-Driven Development: An Empirical Evaluation of Agile Practice. Springer, (Heidelberg, London, New York), DOI 10.1007/978-3-642-04288-1

Madeyski L (2015) reproducer: Reproduce Statistical Analyses and Meta-Analyses. URL `http://madeyski.e-informatyka.pl/reproducible-research`, R package (`http://CRAN.R-project.org/package=reproducer`)

Madeyski L, Jureczko M (2015) Which Process Metrics Can Significantly Improve Defect Prediction Models? An Empirical Study. Software Quality Journal 23(3):393–422, DOI 10.1007/s11219-014-9241-7

Madeyski L, Orzeszyna W, Torkar R, Józala M (2012) Appendix to the paper "Overcoming the Equivalent Mutant Problem: A Systematic Literature Review and a Comparative Experiment of Second Order Mutation". URL `http://madeyski.e-informatyka.pl/download/app/AppendixTSE.pdf`

Madeyski L, Orzeszyna W, Torkar R, Józala M (2014) Overcoming the Equivalent Mutant Problem: A Systematic Literature Review and a Comparative Experiment of Second Order Mutation. IEEE Transactions on Software Engineering 40(1):23–42, DOI 10.1109/TSE.2013.44

Micceri T (1989) The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin 105(1):156–166

Mosteller F, Tukey JW (1977) Data analysis and regression: A second course in statistics. Addison-Wesley

Mudholkar GS, Marchetti CE, Lin CT (2002) Independence characterizations and testing normality against restricted skewness–kurtosis alternatives. Journal of Statistical Planning and Inference 104(2):pp 485–501

Price RM, Bonett DG (2001) Estimating the variance of the sample median. Journal of Statistical Computation and Simulation 68(3):295–305, DOI 10.1080/00949650108812071

R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Ramsey PH (1980) Exact Type 1 Error Rates for Robustness of Student's t Test with Unequal Variances. Journal of Educational and Behavioral Statistics 5(4):337–349, DOI 10.3102/10769986005004337

Razali NM, Wah YB (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics 2(1):21–33

Shadish WR, Cook TD, Campbell DT (2002) Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, MA, USA

Shapiro SS, Wilk M, Chen HJ (1968) A comparative study of various tests for normality. Journal of the American Statistical Association 63(324):1343–1372

Shrout P, Fleiss J (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin 86(2):420–428, DOI 10.1037/0033-2909.86.2.420

Stout DE, Ruble TL (1995) Assessing the practical significance of empirical results in accounting education research: the use of effect size information. Journal Of Accounting Education 13(3):281–298

Tappenden AF, Miller J (2014) Automated cookie collection testing. ACM Transactions on Software Engineering and Methodology 23(1):3:1–3:40, DOI 10.1145/2559936

Tian T, Wilcox R (2007) A comparison of two rank tests for repeated measures designs. Journal of Modern Applied Statistical Methods 6(1):331–335

Urdan TC (2005) Statistics in Plain English, 2nd edn. Routledge, Oxon, UK

Vargha A, Delaney HD (2000) A critique and improvement of the CL common language effect size statistics of McGraw and Wong. Journal of Educational and Behavioral Statistics 25(2):101–132, DOI 10.3102/10769986025002101

Viechtbauer W (2010) Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software 36(3):1–48

Welch BL (1938) The Significance of the Difference Between Two Means when the Population Variances are Unequal. Biometrika 29(3-4):350–362, DOI 10.1093/biomet/29.3-4.350

Whigham PA, Owen C, MacDonell S (2015) A baseline model for software effort estimation. ACM Transactions on Software Engineering and Methodology 24(3):20:1–20:11

Wilcox RR (1998) How many discoveries have been lost by ignoring modern statistical methods? American Psychologist 53(3):300–314

Wilcox RR (2012) Introduction to Robust Estimation & Hypothesis Testing, 3rd edn. Elsevier

Wilcox RR, Keselman HJ (2003) Modern robust data analysis methods: Measures of central tendency. Psychological Methods 8(3):254–274

Yuen KK (1974) The Two-Sample Trimmed t for Unequal Population Variances. Biometrika 61(1):165–170

Zimmerman DW (2000) Statistical Significance Levels of Nonparametric Tests Biased by Heterogeneous Variances of Treatment Groups. Journal of General Psychology 127(4):354–364, DOI 10.1080/00221300009598589

Zimmerman DW, Zumbo BD (1993) Rank transformations and the power of the Student t test and Welch t test for non-normal populations with unequal variances. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale 47(3):523–539